COMPARATIVE GENOMICS OF MICROBIAL SIGNAL TRANSDUCTION

A Dissertation
Presented to
The Academic Faculty

by

Luke Ulrich

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in Bioinformatics

Georgia Institute of Technology
December, 2005

COMPARATIVE GENOMICS OF MICROBIAL SIGNAL TRANSDUCTION

Approved by:

Dr. Igor Zhulin, Advisor
School of Biology
University of Tennessee – Oak Ridge National Laboratory

Dr. Stephen Spiro
School of Biology
Georgia Institute of Technology

Dr. Mark Borodovsky
School of Biology
Georgia Institute of Technology

Dr. Matthew Wolf
College of Computing
Georgia Institute of Technology

Dr. John Kirby
School of Biology
Georgia Institute of Technology

Date Approved: November 10, 2005

I would like to dedicate this dissertation to my father who has constantly encouraged me in my educational pursuits and helped me reach this amazing goal.

ACKNOWLEDGEMENTS

I want to thank Dr. Igor Zhulin for his wisdom, computers, psychotherapy, and Russian jokes. I want to thank my committee members – Dr. Mark Borodovsky, Dr. John Kirby, Dr. Stephen Spiro, and Dr. Matthew Wolf – for their helpful participation. I also want to thank my mother and father for their constant encouragement and support. Thank you Megan for your awesome friendship, the many meals, and help. Thank you Orris for your comic relief and the many lunches we shared as well as your advice on all matters.

TABLE OF CONTENTS

	Page
ACKNOWLEDGEMENTS	iv
LIST OF TABLES	viii
LIST OF FIGURES	ix
LIST OF SYMBOLS AND ABBREVIATIONS	xiv
SUMMARY	xvi
CHAPTER	
1 INTRODUCTION	1
Microbial Signal Transduction	1
Impact and Significance	1
Design and Architecture	3
Comparative Genomics	9
Genomic Data Growth	10
Sequence, Structure, Function Relationship	12
Pairwise Sequence Alignment	14
Multiple Sequence Alignment	19
Comparative Genomics of Microbial Signal Transduction	22
Signaling Domains	23
Objectives	29
2 GENERAL MATERIALS AND METHODS	30
Databases	30
Sequence Databases	31
Structural Databases	34

Domain Databases 34

Tools 36

Sequence Similarity Search Tools 36

Multiple Sequence Alignment and Phylogenetics 39

Secondary Structure Prediction 44

Domain Architecture Prediction 47

Visualization 48

3 BIOINFORMATICS PLATFORM AND THE MIST DATABASE 50

Introduction 50

Hardware and Software 50

The MiST Database 55

High-throughput Identification of Signal Transduction Proteins 59

Exploratory Knowledge Environment 66

4 ONE-COMPONENT REGULATORS DOMINATE SIGNAL TRANSDUCTION IN PROKARYOTES 77

Introduction 77

What is a One-component System? 78

Detection of Signal Transduction Proteins in Sequenced Genomes 80

One-component Versus Two-component Systems: a Survey of Bacterial and Archaeal Genomes 82

Genome Size, Lifestyle, and Environment Contribute to the Complexity of Signal Transduction 85

One-component Systems as the Primordial Form of Prokaryotic Signal Transduction 87

Conclusions 88

Acknowledgements 89

5 FOUR-HELIX BUNDLE: A UBIQUITOUS SENSORY MODULE IN PROKARYOTIC SIGNAL TRANSDUCTION 90

Abstract 90

Introduction 91

Methods 92

Results and Discussion 95

6 RESOLVING THE FUNCTION OF CHEMOTAXIS PAS DOMAINS THROUGH PROTEIN SEQUENCE ANALYSIS 103

Abstract 103

Introduction 104

Methods 106

Results and Discussion 109

Acknowledgements 117

7 CONCLUSION 118

APPENDIX A: Supplementary Information for Chapter 4 119

APPENDIX B: Supplementary Information for Chapter 5 126

APPENDIX C: Supplementary Information for Chapter 6 156

APPENDIX D: Publications 168

REFERENCES 169

LIST OF TABLES

		Page
Table 1.1:	Experimentally characterized signal transduction pathways.	3
Table 1.2:	Sensor histidine kinases from the genome of *Campylobacter jejuni*.	25
Table 3.1:	Pfam and SMART domains used for identifying signal transduction proteins.	64
Table 6.1:	E-values and scores of the last true positive and first true negative from searches of the PAS_Aer and PAS_Che profiles against the non-redundant database (4 January 2005). Scores are given in the parenthesis following the E-value.	108
Table 6.2:	Identification of PAS domains in methyl-accepting chemotaxis proteins by HMM domain profiles.	109
Table A.1:	Domains and domain categories used to identify signal transduction systems.	119
Table A.2:	Distribution of two-component and one-component systems in prokaryotic genomes.	120
Table A.3:	Genomic distribution of input and output domains in archaea and bacteria.	124

LIST OF FIGURES

Page

Figure 1.1: Prototypical, bacterial two-component system. In response to a particular stimulus detected by its input domain, the sensor kinase undergoes an ATP-dependent autophosphorylation at a conserved histidine residue in its transmitter domain. The cognate response regulator then catalyzes the transfer of this phosphoryl group to a conserved aspartate residue on its receiver domain, which effects an adaptive response via its output domain. 4

Figure 1.2: The EnvZ-OmpR two-component system of *E. coli*. EnvZ senses osmolarity changes and transmits this signal to its cognate response regulator, OmpR, via a phosphotransfer reaction. In response, OmpR regulates the expression of the *ompF* and *ompC* genes, which encode the outer membrane porins, OmpF and OmpC, respectively. 5

Figure 1.3: The BvgS-BvgA phosphorelay of *B. parapertussis*. BvgS senses an unknown signal via its PBP and PAS input domains, which results in the phosphorylation of BvgA via a His-Asp-His-Asp phosphorelay reaction. BvgA responds by regulating the expression of biofilm related genes. Vertical, blue bars represent transmembrane regions. 6

Figure 1.4: The chemotaxis system of *E. coli*. An MCP detects a chemical stimulus and transmits this information to the CheA histidine kinase via protein-protein interactions and the adaptor protein, CheW. CheA undergoes an autophosphorylation reaction at a conserved histidine residue, and subsequently releases the phosphoryl group to the specialized response regulator, CheY. In response, CheY regulates the flagella direction by directly binding to the motor apparatus. CheR and CheB reset the MCP enabling it to sense future changes in the chemical concentration. The phosphatase, CheZ, resets CheY. Vertical, blue bars represent transmembrane regions. 8

Figure 1.5: Exponential growth of NCBI's GenBank over the past decade. In 2005, GenBank exceeded 100 gigabases of nucleotide data. 11

Figure 1.6: Number of completely sequenced genomes over the past decade. 11

Figure 1.7: Sequence, structure, function relationship and the role of comparative genomics. 13

Figure 3.1: Structure of the Microbial Signal Transduction database, MiST. Table names are in bold type and lines indicate the various relationships between tables. Values above the lines signify the cardinality constraints for each relationship. The labels 'pk' and 'fk' refer to primary key and foreign key, respectively. In some cases, a single column is both a primary and foreign key, in which case only 'pk' is listed. The color of each table denotes a particular database section: light blue – primary genomic data; purple – derived data; orange – results; and green – management. 58

Figure 3.2: High-throughput identification of signal transduction proteins. 59

Figure 3.3: Web interface to PSI-BLAST. Input spaces are provided for the query sequence, database to search, and other various options. Users may also load previous searches using this page. 67

Figure 3.4: Output of results from a web-based PSI-BLAST search. Descriptive information is given about the search including the BLAST version, database size, and number of hits. A graphical output displays hits to the query sequence as horizontal lines and the color of each line indicates its score. The search may be continued additional iterations or the results from this search used in another analysis (e.g. view the domain architectures for the selected sequences). Significant BLAST hits and information about each pairwise alignment are displayed below the graphical overview. 68

Figure 3.5: Web interface to the Consensus tool. Users upload ClustalW alignments or input the alignment directly into the provided text area. Checkboxes specify the consensus levels to produce. 69

Figure 3.6: Output of results from the web-based Consensus tool. The alignment is redisplayed with the consensus above and below. 69

Figure 3.7: Web interface to the Alignment Shader tool. Users upload ClustalW alignments or input the alignment directly into the provided text area. The user then describes the groups of sequences within the alignment, what color to shade them, and at what similarity threshold a column must surpass for it to be shaded. 70

Figure 3.8: Output of results from the web-based Alignment Shader tool. 70

Figure 3.9: Entry web page to the MiST database. Available genomes are listed according to their taxonomy. Clicking the individual checkboxes beside each organism name selects them for further investigation. Users may also search by GI number or MiST identifiers for a specific protein. 72

Figure 3.10: Analysis selection web page for the MiST database. Each of the previously selected organisms may be searched by domain, description, GI, or MiST identifiers. Clicking on an individual organism name displays various information for that particular organism. 73

Figure 3.11: Organism specific page for the MiST database containing descriptive information about the genome, signal transduction profile, querying options, and lists of one- and two-component proteins found in this organism. 74

Figure 3.12: Protein and gene web page for the MiST database. Basic annotation and sequence data for a protein and its corresponding gene are displayed along with the predicted domain architecture and genome neighborhood. 76

Figure 4.1: Two-component and one-component signal transduction. (a) A prototypical two-component signal transduction system contains input (colored red) and output (colored yellow) domains in two different proteins that communicate via a His-Asp phosphotransfer. A one-component system is a protein that contains input and output domains but lacks His-Asp phosphotransfer domains (colored gray and white). (b) Examples of two-component and one-component systems that utilize the same type of input and output domains. Experimentally studied proteins are identified by name, while proteins predicted from genome sequences are identified by their GenBank ID number. 80

Figure 4.2: Distribution of input and output domains in bacterial and archaeal signal transduction systems. The counts of the twenty-five most abundant input and output domains in bacterial and archaeal one-component and two-component systems are shown. Domain nomenclature is from the curated Pfam-A database (see Appendix A for detailed information). 84

Figure 4.3: Dependence of the number of one-component and two-component signal transduction systems on the genome size. The plot is in a double logarithmic scale. One hundred forty-five genomes were ranked by size and split into 16 size classes. Each point indicates the average number of genes for one-component or two-component signal-transduction systems in the respective class. 86

Figure 5.1: Overview of the VISSA process. For each sequence in a multiple alignment, its secondary structure is predicted and both the alignment and structural information stored in an XML document. This XML data is then visualized by coloring/shading the amino acids that correspond to the predicted secondary structure elements. 94

Figure 5.2: Alignment of representative members of the 4HB_MCP domain. (**A**) An alignment of thirty members of the 4HB_MCP domain from the seed alignment. Conserved residues and their positions are colored with the ClustalX scheme using Jalview (Clamp, et al., 2004): orange – glycine (G), yellow – proline (P), blue – small and hydrophobic amino acids (A, V, L, I, M, F, W), green – hydroxyl and amine amino acids (S, T, N, Q), red – charged amino acids (D, E, R, K), cyan – histidine (H) and tyrosine (Y). (**B**) An alignment of the same thirty members illustrating the VISSA visualization. Regions containing predicted alpha helices and beta sheets have the background shaded red and blue, respectively. The intensity of each shade is directly proportional to the confidence of a given prediction – a darker intensity representing a higher confidence. The identifier for each sequence consists of a species abbreviation, the GenBank identifier, and the coordinates of the sequence. 96

Figure 5.3: Visualization of conserved residues in the 4HB_MCP domain. (**A**) A sequence logo generated for the multiple alignment of 282 domain sequences obtained in the PSI-BLAST search initiated with the ligand-binding domain of the *E. coli* Tar chemoreceptor. The secondary structure of Tar is shown above the logo. (**B**) Conserved tyrosine residues in helices 2 and 3 interact to maintain packing of the four-helix bundle. 99

Figure 5.4: A schematic view of the domain architecture of 4HB_MCP containing proteins. Domain nomenclature is according to Pfam (Bateman, et al., 2004). Protein GenBank identifiers and the species abbreviations are shown. 101

Figure 6.1: Multiple sequence alignment with the VISSA visualization of the three subfamilies of PAS domains found in methyl-accepting chemotaxis proteins. Thirty representative PAS domains from each subfamily are shown (for the full alignment of 274 sequences, see Appendix C, Figure C.2). The multiple alignment was constructed using the PCMA (Pei, et al., 2003) and ClustalW (Thompson, et al., 1994) programs, visualized using the VISSA protocol (Ulrich and Zhulin, 2005). Regions containing predicted alpha helices and beta strands have the background shaded in red and blue, respectively. The shading intensity is directly proportional to the confidence of a given prediction – a darker intensity representing a higher confidence. The PAS core, helical connector, and beta scaffold structural regions are delineated above the alignment. Structural elements, as defined by Gong *et al*. (1998), are listed below the alignment. 111

Figure 6.2: Neighbor-joining tree built from a multiple sequence alignment of 274 PAS domains showing three distinct subfamilies. 112

Figure 6.3: Sequence logos for three PAS subfamilies. Highlighted positions are strongly conserved (BLOSUM consensus at least 85% or information content greater than 3 bits). Residues unique to a given subfamily are designated with a triangle (▲). Red arrows indicate residues critical for binding FAD by the PAS domain in *E. coli* Aer and black arrows indicate other residues that are critical for Aer function (Repik, et al., 2000). 112

Figure B.1: Unedited seed 4HB_MCP alignment with VISSA visualization. 126

Figure B.2: Edited seed 4HB_MCP alignment with VISSA visualization. 135

Figure B.3: Complete 4HB_MCP alignment with VISSA visualization. 141

Figure C.1: Unedited alignment of chemotaxis PAS domains constructed using PCMA and ClustalW, and visualized with VISSA. 156

Figure C.2: Alignment of chemotaxis PAS domains after manual editing and visualized using VISSA. 162

LIST OF SYMBOLS AND ABBREVIATIONS

BLAST	Basic Local Alignment Search Tool
BLOSUM	Blocks Amino Acid Substitution Matrices
CGI	Common Gateway Interface
COG	Cluster of Orthologous Groups
DBI	Database Interface
DNA	Deoxyribonucleic Acid
DP	Dynamic programming
DTD	Document Type Definition
EST	Expressed sequence tag
FLOP	Floating-point operation
FTP	File Transfer Protocol
GB	Gigabyte
GHz	Gigahertz
GI	GenBank identifier
HMM	Hidden Markov model
HTH	Helix-turn-helix
HTML	Hypertext Markup Language
k-mer	A contiguous subsequence of length k
LWP	Library for the World-wide-web for Perl
MCP	Methyl-accepting chemotaxis protein
MEGA	Molecular Evolutionary Genetics Analysis
MiST	Microbial Signal Transduction
NCBI	National Center for Biotechnology Information

NR	Non-redundant database
OS	Operating system
PAM	Point Accepted Mutation
PCMA	Profile Consistency Multiple Sequence Alignment
PDB	Protein Data Bank
Pfam	Protein families
PSI-BLAST	Position Specific Iterative Basic Local Alignment Search Tool
PSSM	Position Specific Scoring Matrix
RAID	Redundant Array of Inexpensive Disks
RAM	Random access memory
RDBMS	Relational Database Management System
RefSeq	Reference Sequence database
SGE	Sun Grid Engine
SMART	Simple Modular Architecture Research Tool
SQL	Standard Query Language
T-Coffee	Tree-based Consistency Objective Function for Alignment Evaluation
VMD	Visual Molecular Dynamics
XML	eXtensible Markup Language

SUMMARY

The goal of this research is to translate genomic sequence data into high-quality biological knowledge on microbial signal transduction. Signal transduction pathways control important cellular functions in all organisms including biosynthesis, catabolism, chemotaxis, development, virulence, host-recognition, and antibiotic resistance. Genome sequencing projects have generated a vast amount of data that has not been analyzed in detail either computationally or experimentally. Up to 40% of genes in most microbial genomes lack assigned biological functions. In the foreseeable future, this gap between the number of sequenced genes and the extent of their functional characterization is expected to broaden further. This creates an urgent need for improved computational prediction of biological function. Signal transduction is one of the most problematic areas for current automated genome annotation protocols, because of the high sequence variability of input and output domains and the mosaic domain architecture of signal transduction proteins. Furthermore, microbiologists studying signal transduction do not actively participate in the validation of computational predictions and the annotation process. This research aims at closing the existing gap between genomic and experimental data in the area of signal transduction in the simplest organisms - prokaryotes.

High-throughput genome processing, sophisticated protein sequence analysis, programming, and information management were used to achieve two major advances in the comparative genomics of microbial signal transduction:

Research Advance 1: *A bioinformatics platform and the Microbial Signal Transduction (MiST) database were designed and implemented*. An integrated and robust bioinformatics platform and the Microbial Signal Transduction database (MiST) were developed, which facilitated the genome-wide analysis of bacterial signal transduction. This platform was successfully used for the high-throughput identification and classification of over 55,000 signal transduction proteins in more than 300 archaeal and bacterial organisms.

Research Advance 2: *One-component systems dominate signal transduction in prokaryotes*. Two-component systems that link environmental signals to cellular responses are viewed as the primary mode of microbial signal transduction. A comprehensive review of the signal transduction systems encoded in prokaryotic genomes revealed that contrary to this view, the majority of signal transduction systems are one-component systems – a single protein containing both input and output domains but lacking phosphotransfer domains typical of two-component systems. One-component systems are more widely distributed among bacteria and archaea and display a greater diversity of domains than two-component systems.

Additionally, in-depth bioinformatic analyses were performed that further characterized the function of two, input, signaling domains. Transmembrane chemoreceptors in *Escherichia coli* utilize ligand-binding domains for detecting various external signals. Current domain models for these important sensory modules were constructed from unrelated proteins and fail to detect the majority of domain homologs. A more accurate model, the four-helix up and down bundle, was described that represents

a large domain family with representatives from all major classes of prokaryotic signal transduction including histidine kinases, di-guanylate cyclases, and chemotaxis receptors. PAS domains constitute a widespread superfamily of sensory modules that primarily sense light, oxygen, small ligands, and redox potential. Current computational techniques identify PAS domains in thousands of protein sequences; however, they fail to predict the specific function of a given PAS domain. Two hundred seventy-four PAS domains found in association with chemotaxis receptors were classified into distinct subfamilies. Based on experimental evidence, members belonging to the PAS_Aer subfamily are predicted to function as redox sensors. Members of the PAS_Che subfamily are predicted to bind an unknown ligand and represent a good target for experimental studies. This analysis of chemotaxis PAS domains demonstrates that standard bioinformatics approaches may be applied to multifunctional domains in order to further resolve their function.

CHAPTER 1

INTRODUCTION

Microbial Signal Transduction

For an organism to survive and adapt to changing conditions, it must be able to sense and respond appropriately to environmental signals. Any disruption in an organism's ability to interact with its environment usually produces fatal results or an aberrant and stunted lifestyle. For example, a bacterium ceases to function optimally if it is unable to locate a sustainable energy source or detect changes in the surrounding osmolarity. To prevent useless expenditure of energy, pathogens must recognize a suitable host system before releasing virulent toxins. To combat such pathogens, the host must be able to sense these foreign invaders and then respond with the appropriate immune response. The processing of environmental stimuli directly or indirectly influences the majority of cellular activities in all organisms ranging from bacteria to humans. Therefore elucidating the components and mechanisms of how an organism senses and responds to environmental signals enables a much deeper understanding of the organism and how it behaves, interacts, and adapts to its particular environmental niche.

Impact and Significance

The study of microbial signal transduction includes understanding bacterial systems and pathways responsible for sensing and responding to environmental cues. These systems primarily consist of one or more proteins that behave as sensors and/or regulators that detect a particular stimulus and initiate an adaptive cellular response. Such signaling pathways control the majority of cellular functions including transport,

osmoregulation, respiration, metabolism, biosynthesis, chemotaxis, and development (Stock, et al., 2000). Table 1.1 contains a list of several experimentally characterized systems. Furthermore, signal transduction systems control a significant portion of the genome. For example, the ArcB/ArcA signal transduction system in *Escherichia coli* regulates the expression of more than thirty operons in response to changes in redox conditions (Georgellis, et al., 2001). The NtrB/NtrC system controls up to 2% of the *E. coli* genome in response to nitrogen availability (Zimmer, et al., 2000). In addition, signal transduction systems regulate virulence, host recognition, and antibiotic resistance in important human pathogens, such as *Staphylococcus aureus* (Novick and Jiang, 2003), *Streptococcus pneumonia* (Blue and Mitchell, 2003), *Streptococcus pyogenes* (Biswas and Scott, 2003), *Vibrio cholerae* (Krukonis and DiRita, 2003), *Mycobacterium tuberculosis* (Zahrt and Deretic, 2001), and many others. Signal transduction systems have been reported in the genomes of bacteria that are used in biological warfare, namely *Bacillus anthracis* (Read, et al., 2003) and *Yersinia pestis* (Deng, et al., 2002). Proteins involved in these systems of various pathogenic bacteria are attractive targets for antimicrobial drug design (Matsushita and Janda, 2002; Stephenson and Hoch, 2002). Because of its critical role in bacteria, signal transduction heavily impacts basic research and expands into important industries including medicine, agriculture, biological warfare and bioterrorism, and bioremediation.

Table 1.1 Experimentally characterized signal transduction pathways.

	Signaling pathway	Genes	Reference
Transport	Citrate-update	*citA, citB*	(Reinelt, et al., 2003)
		bctD, bctE	(Antoine, et al., 2005)
	C_4-dicarboxylates	*dctB, dctD*	(Reid and Poole, 1998)
	Iron	*ritR*	(Ulijasz, et al., 2004)
	Potassium	*kdpD, kdpE*	(Walderhaug, et al., 1992)
Osmo-regulation	*Outer membrane porins*	*envZ, ompR*	(Cai and Inouye, 2002)
		cpxA, cpxR	(Batchelor, et al., 2005)
Respiration	Aerobic	*arcB, arcA*	(Georgellis, et al., 2001)
	Redox-control	*regB, regA*	(Elsen, et al., 2004)
Metabolism	Carbon	*uhpB, uhpA*	(Island and Kadner, 1993)
	Nitrogen	*ntrB, ntrC*	(Zimmer, et al., 2000)
	Phosphate	*phoR, phoB*	(Danhorn, et al., 2004)
Chemotaxis	General	*cheA, cheY, cheB, cheR, cheZ*	(Armitage, 1999)
Development	Sporulation	*kinA, kinB, spo0A, spo0B, spo0F, rapA, rapB*	(Stephenson and Hoch, 2002)
Virulence	Pneumococcal	*pnpR, pnpS*	(McCluskey, et al., 2004)
	Biofilm development	*bvgA, bvgS*	(Mishra, et al., 2005)
Host recognition	Agrobacterium pathogenesis	*virA, virG*	(Cho and Winans, 2005)
Antibiotic resistance	Vancomycin	*vanR, vanS*	(Haldimann, et al., 1997)
	Bacitracin	*bceR, bceS*	(Ohki, et al., 2003)
	Beta-lactams	*croR, croS*	(Comenge, et al., 2003)

Design and Architecture

Two-component Systems

Common to all bacterial signal transduction systems is the detection of a signal (input) and coupling this with a cellular response (output). The most widely recognized

signaling systems are so-called two-component systems that utilize protein phosphorylation as a fundamental strategy for signaling (Hoch and Silhavy, 1995; Inouye and Dutta, 2003; Stock, et al., 2000). The prototypical two-component system consists of two proteins, a sensor kinase and a response regulator (Figure 1.1). The sensor kinase detects an environmental signal via its input domain(s). This results in an ATP-dependent autophosphorylation of a conserved histidine residue within the transmitter domain. The response regulator then catalyzes the transfer of the phosphoryl group to a conserved aspartate residue within its receiver domain. Phosphorylation of the receiver domain activates the output domain(s), which effects a particular adaptive response, usually regulation of transcription (Parkinson and Kofoid, 1992).

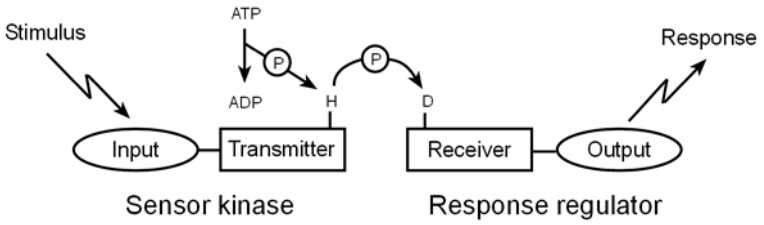


Figure 1.1 Prototypical, bacterial two-component system. In response to a particular stimulus detected by its input domain, the sensor kinase undergoes an ATP-dependent autophosphorylation at a conserved histidine residue in its transmitter domain. The cognate response regulator then catalyzes the transfer of this phosphoryl group to a conserved aspartate residue on its receiver domain, which effects an adaptive response via its output domain.

One of the best-studied systems, the EnvZ-OmpR two-component system of *E. coli* is an example of such a classical signal transduction pathway (Figure 1.2). This pathway regulates the expression of two outer membrane porins, OmpF and OmpC, in response to changes in the osmolarity of the surrounding medium. OmpF and OmpC form channels in the outer membrane that enable the diffusion of small hydrophilic molecules and thus maintain the osmotic pressure of the cell. EnvZ is a membrane bound histidine kinase that senses osmolarity changes via its periplasmic sensory domain. This event triggers the phosphorylation of the conserved histidine in the conserved kinase core of EnvZ. Consequently, this phosphoryl group is transferred to the conserved aspartate residue of the N-terminal, receiver domain of OmpR. This in turn activates the C-terminal, DNA-binding domain of OmpR, which differentially regulates the *ompF* and *ompC* genes and thus appropriately modulates the cell's osmotic pressure (Cai and Inouye, 2002).

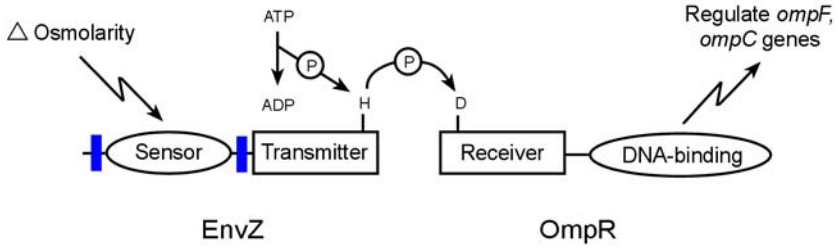


Figure 1.2 The EnvZ-OmpR two-component system of *E. coli*. EnvZ senses osmolarity changes and transmits this signal to its cognate response regulator, OmpR, via a phosphotransfer reaction. In response, OmpR regulates the expression of the *ompF* and *ompC* genes, which encode the outer membrane porins, OmpF and OmpC, respectively. Vertical, blue bars represent transmembrane regions.

Phosphorelays

Phosphorelays extend the basic two-component theme with the inclusion of additional histidine (Hpt) and aspartate containing domains (receiver), which involve multiple phosphotransfer reactions between two or more proteins (Appleby, et al., 1996; Hoch, 2000). The BvgAS two-component system of *Bordetella parapertussis* uses such a phosphorelay in regulating biofilm development (Figure 1.3). In response to an unknown signal (detected by the PBP and PAS input domains), BvgS autophosphorylates at the conserved histidine residue within its transmitter domain. This phosphoryl group then cascades along a His-Asp-His-Asp phosphorelay (i.e. transmitter-receiver-Hpt-receiver), which results in the binding of DNA by the response regulator, BvgA, and regulation of *bvg* related genes (Mishra, et al., 2005). Phosphorelays provide more points of control for fine-tuning the regulation of a system than simpler two-component systems.

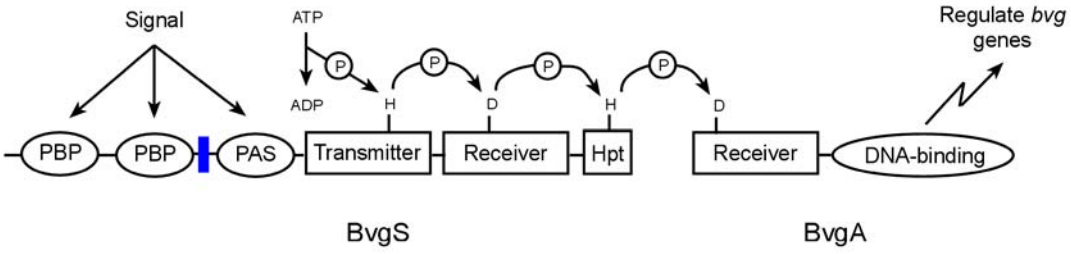


Figure 1.3 The BvgS-BvgA phosphorelay of *B. parapertussis*. BvgS senses an unknown signal via its PBP and PAS input domains, which results in the phosphorylation of BvgA via a His-Asp-His-Asp phosphorelay reaction. BvgA responds by regulating the expression of biofilm related genes. Vertical, blue bars represent transmembrane regions.

Chemotaxis

Chemotaxis is the ability of motile organisms to navigate towards or away from a chemical stimulus and is controlled by one of the most sophisticated signal transduction pathways in microbes (Parkinson, 1993). While there are variations in the chemotaxis pathway(s) between organisms, the *E. coli* chemotaxis pathway is the best-understood and studied system (Figure 1.4). Specialized transmembrane receptors, also known as methyl-accepting chemotaxis proteins (MCPs), respond to changing chemical concentrations and in turn interact with a cytoplasmic histidine kinase, CheA, via the adapter protein, CheW (Boukhvalova, et al., 2002; Falke and Hazelbauer, 2001; Parkinson, 1993). CheA autophosphorylates at its conserved histidine residue and then transfers this phosphoryl group to the receiver domain of its cognate response regulator, CheY. CheY lacks an output domain typical of most response regulators, and instead its receiver domain directly interacts with the flagellar motor to adjust the direction of the flagellar apparatus (Schuster, et al., 2001). CheR and CheB chemically modify the MCP using methyl groups to reset its ability to sense differing chemical concentrations (Djordjevic and Stock, 1998). Finally, CheZ dephosphorylates CheY, restoring its ability to participate with CheA (Zhao, et al., 2002).

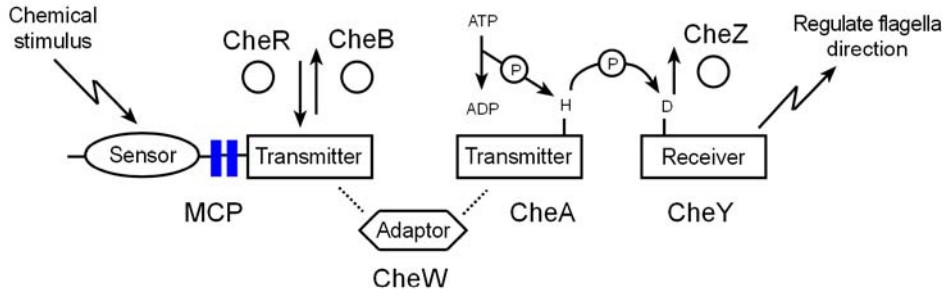


Figure 1.4 The chemotaxis system of *E. coli*. An MCP detects a chemical stimulus and transmits this information to the CheA histidine kinase via protein-protein interactions and the adaptor protein, CheW. CheA undergoes an autophosphorylation reaction at a conserved histidine residue, and subsequently releases the phosphoryl group to the specialized response regulator, CheY. In response, CheY regulates the flagella direction by directly binding to the motor apparatus. CheR and CheB reset the MCP enabling it to sense future changes in the chemical concentration. The phosphatase, CheZ, resets CheY. Vertical, blue bars represent transmembrane regions.

Proteolytic Signaling Pathways

In addition to the prevalent and widespread recognition of two-component systems, alternative signaling pathways have been described. Staphylococcal species employ a proteolytic signaling pathway to confer resistance to β-lactam based antibiotics. The sensor protein, BlaR1, undergoes an autocatalytic cleavage upon binding of β-lactam antibiotics to its penicillin-binding domain. One of the resulting fragments contributes to cleaving the DNA-binding protein, BlaI, which then dissociates from the *blaZ* gene locus. Consequently, the β-lactamase enzymes are produced resulting in resistance to such antibiotics (Zhang, et al., 2001).

In general, signal transduction systems are highly modular. Variable input and output domains typically associate via conserved transmitter and receiver domains.

Sensor kinases are typically membrane bound with one or more N-terminal sensory domains for detecting a particular stimulus, and a C-terminal transmitter domain containing a conserved kinase core for downstream signaling. Due to the broad range of signals they must sense, input domains comprise a very diverse, highly variable group.

Response regulators are characterized by an N-terminal receiver domain with a conserved aspartate residue used in phosphotransfer reactions and normally contain one or more output domains. These effector domains are more conserved than input domains and tend to have fewer functional roles. The most commonly found output domains are DNA-binding helix-turn-helix (HTH) domains because the predominant adaptive response from response regulators is control of gene expression, which is mediated by interacting with DNA (Parkinson and Kofoid, 1992). Other output responses include RNA anti-termination, regulation of enzymatic activity via adenylate and di-guanylate cyclases, c-di-GMP-phosphodiesterases, phosphohydrolases, and other related domains (Aravind and Koonin, 1998; Chai and Stewart, 1998; Chan, et al., 2004; Jenal, 2004).

In summary, signal transduction represents an incredibly significant cellular activity impacting numerous fields of study and industries. In addition to the functional mechanisms of sensing and responding, other studies document novel signal transduction pathways and the interacting proteins in these pathways (Eguchi and Utsumi, 2005), cross-species variation (von Mering, et al., 2003), structural and biochemical signaling mechanisms (Nioche, et al., 2004; Pappas, et al., 2004), and their environmental implications (Balazsi, et al., 2005).

Comparative Genomics

The nature of experimental characterization of a given gene or protein demands considerable time and resources. Consequently, high-throughput genome sequencing projects uncover novel gene and protein sequences much faster than they may be experimentally analyzed. This underscores the need for computational systems capable of

accurately mapping functional information from known, characterized systems onto the genomic sequence space. This is precisely the goal of comparative genomics, which focuses on the comparison of sequence data from multiple genomes to functionally annotate genes and proteins.

Growth of Genomic Data

Over the past decade, advances in genetics, molecular biology, DNA sequencing techniques and instrumentation have made possible the rapid determination of an organism's DNA – also known as its genome. As a result, numerous high-throughput sequencing initiatives have generated and are continuing to generate, massive quantities of genomic data from all kingdoms of life. This genomic data continues to exponentially increase as evidenced by the growth of GenBank (Benson, et al., 2005), a public repository of nucleotide sequences, which recently exceeded 100 gigabases of nucleotides (Figure 1.5). As of October 2005, genome-sequencing projects have resulted in more than 300 completely sequenced genomes (Figure 1.6) with more than 1600 other genomic projects underway (Bernal, et al., 2001). It is expected that this rate of genomic data production will continue to accelerate with the development of massively parallel sequencing methods that are able to sequence up to twenty-five megabases of DNA in a single four hour period (Rogers and Venter, 2005).

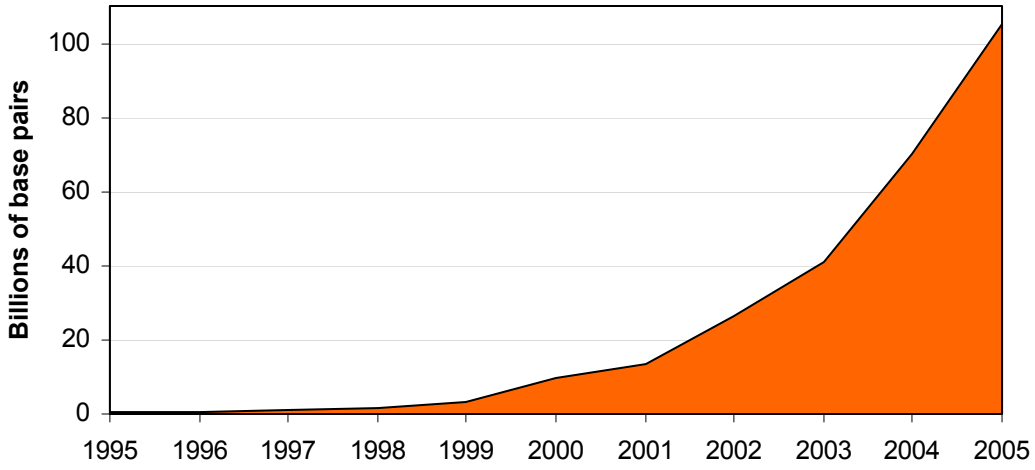


Figure 1.5 Exponential growth of NCBI's GenBank over the past decade. In 2005, GenBank exceeded 100 gigabases of nucleotide data.

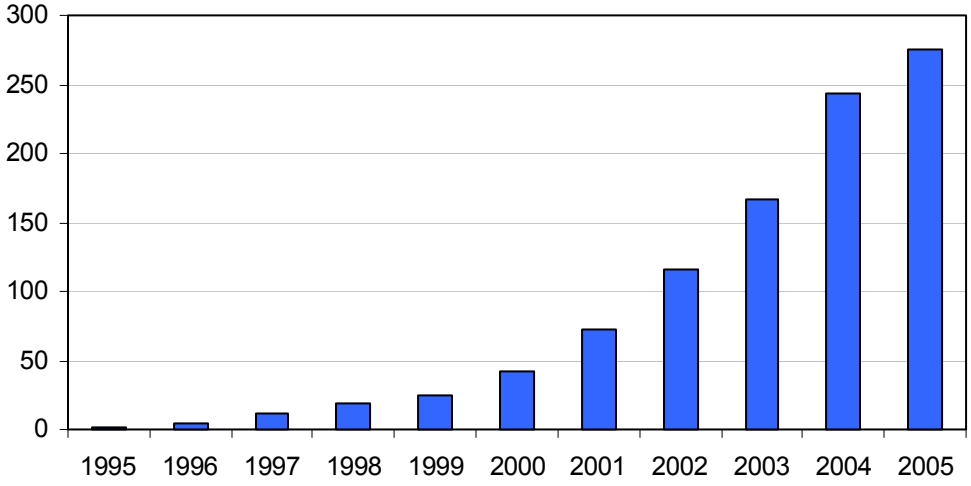


Figure 1.6 Number of completely sequenced genomes over the past decade.

This mountain of genomic data has driven the development of computational tools and methodologies for decoding the stockpiles of DNA into meaningful information. One of the most fundamental and useful tasks entails identifying the genes within a genome. For microbial genomes, computational gene finding tools such as GeneMark (Besemer, et al., 2001) and Glimmer (Delcher, et al., 1999) readily identify 99% of all protein-coding genes. Eukaryotic gene structure is much more complex than prokaryotes due to their exon/intron organization and dissemination across much larger genomes. Therefore, accurate eukaryotic gene delineation remains an active research problem, and sensitive and specific tools are currently being investigated.

Deriving the proteome, the set of all proteins encoded by the genes of a genome, involves determining the protein sequence by translating each gene sequence using the genetic code. This is a trivial matter for microbial genes, but the alternative splicing of eukaryotic genes complicates the accurate prediction of eukaryotic protein sequences. Correctly decoding the gene and protein sequences from microbial DNA marked a major achievement for genomics research, yet was followed by the challenging problem of accurately predicting the function of each putative protein – the goal of comparative genomics.

Sequence, Structure, Function Relationship

Comparative genomics is the science of characterizing the functions of genes and (their derivative) proteins by comparing sequence data from multiple genomes. This process strongly relies on the concept that a protein sequence determines its structure. This protein arrangement presumably determines its function (Figure 1.7). A protein sequence symbolizes the primary structure of a protein as a linear series of amino acids. While this is a convenient representation, the one-dimensional portrayal of a protein

sequence is an artificial construct that poorly models a protein's state *in vivo*. In a biological system, the natural molecular forces act on the linear chain of amino acids, folding the polypeptide chain into a complex 3D structure. The actual shape and spatial configuration of this folded protein determines its specific biological function. The 3D structure of a protein may be determined using techniques such as X-ray crystallography and nuclear magnetic resonance; however, while such methods provide great insight into the actual mechanism and function of a protein, they are quite expensive in terms of equipment, sample preparation, personnel, and finances. Much work has been targeted towards the *ab initio* prediction of a protein's structure from its sequence, but this remains an active research problem and is beyond the scope of this dissertation. Comparative genomics operates under the assumption that two similar (or identical) sequences are likely to produce similar structures with similar biological functions. Presumably, two functionally similar proteins will share equivalent positional and functional residues at the sequence level. Consequently, similar sequences may be assigned a putative function based on comparisons to other similar proteins without knowing the protein's actual structure.

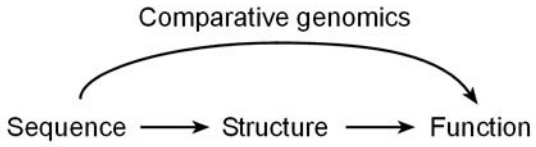


Figure 1.7 Sequence, structure, function relationship and the role of comparative genomics.

Pairwise Sequence Alignment

The alignment of related sequences is fundamental to nearly every comparative genomic study. Sequence alignment is the procedure of searching for common character patterns between sequences in order to reveal their similarity and mutual features. This process involves placing the sequences in rows and shifting the characters (without changing their order) in each sequence such that identical or similar characters are aligned in the same column. Gaps (represented by dashes) may be inserted in order to increase the number of matching characters. An optimal alignment has the maximum possible number of identical and similar characters aligned in each column. A pairwise alignment consists of two sequences aligned to each other, and a multiple alignment comprises three or more simultaneously aligned sequences.

Sequence alignments may be either global or local. A global alignment involves aligning the entire sequences from end to end and is most applicable for highly similar sequences of approximately the same length. A local alignment consists of relatively short subsequences aligned to each other. Local alignments are more appropriate for sequences that share a conserved domain or other functional region that does not extend the entire length of the sequence. In general, most globular proteins consist of multiple functional domains, which occur in a variety of combinations with other domains. Therefore, similarity many times is restricted to relatively short subsequences and protein sequence analysis relies heavily on local alignments at the domain level. This is especially true for analyzing signal transduction proteins, which demonstrate substantial sequence diversity and domain shuffling in their input and output domains.

Scoring

Critical to the alignment process is an appropriate scoring system that numerically defines sequence similarity. Ideally, this will score similar residues higher than dissimilar

residues and a higher score will reflect a more similar biological function. Unfortunately, there is no analytical means for determining a scoring scheme that precisely models all biological systems. Thus, most sequence similarity tools rely on empirically derived schemes for defining similarity scores between every pair of residues. One of the simplest approaches assigns scores based on identical residues – favoring identities and penalizing non-identical residues. While this might suffice for aligning closely related protein sequences, it is well known that sequences with low-identity may have significant similarity and function the same (Koonin and Galperin, 2003). A far superior solution is the use of substitution matrices. A substitution matrix is a 2D probability matrix containing values for every possible pair of residues and numerically represents the odds for which one residue is likely to be substituted by itself or another residue. For mathematical convenience, these odds scores are typically converted into logarithms and thus called log-odds scores. Several different substitution matrices have been derived including ones that attempt to model the chemical, functional, charge, and structural properties of the various amino acids (Karlin and Ghandour, 1985) or based solely on structural similarities (Feng, et al., 1985); however, the two most popular matrices are the PAM (Dayhoff, et al., 1978) and BLOSUM (Henikoff and Henikoff, 1992) matrices, which are based on observed counts of amino acid changes within closely related sequences.

Substitution Matrices

One of the first substitution matrices to take on widespread use was Dayhoff's PAM (Point Accepted Mutation) matrices (Dayhoff, et al., 1978). These matrices are based on 1572 substitutions found within 71 groups of protein sequences that were at least 85% similar with the intent to capture the relative mutability of each amino acid over a given period of evolutionary time. The underlying assumption when building these matrices is that the rate of mutability may be extrapolated to produce additional matrices

that more accurately weight the likelihood of an amino acid change over longer evolutionary times. The PAM-1 matrix is equivalent to 1% divergence of a protein or one amino acid difference per 100 residues. Because the mutability of each amino acid is assumed to be independent and modeled with a Markov chain, the PAM-1 matrix may be multiplied by itself N times to produce a PAM-N matrix. For example, the PAM-250 would be the PAM-1 matrix extrapolated 250 times and would score pairs of residues based on 250 substitutions per 100 residues. Despite this large number of substitutions, sequences at this level of divergence are still about 20% similar.

Unlike the PAM matrices, the BLOSUM (Blocks Amino Acid Substitution Matrices) matrices are not based on an explicit evolutionary model, but are derived from substitution counts observed in approximately 2000 conserved amino acid patterns known as blocks. These blocks were taken from the BLOCKS database (Henikoff and Henikoff, 1991), which at the time represented more than 500 distinct protein families classified by their biochemical function. Each block is an ungapped alignment of an amino acid pattern shared by members of a particular protein family. The number of amino acid pairs was counted from each block and converted into log-odds substitution scores. Several distinct matrices were built from blocks with differing levels of identity. For example, one of the most popular matrices, BLOSUM62, was constructed from sequences that were at least 62% identical (Henikoff and Henikoff, 1992).

Methods

The primary method for constructing pairwise alignments ordinarily includes some form of dynamic programming (DP). DP was originally used for making global alignments (Needleman and Wunsch, 1970) and later modified to produce local alignments (Smith and Waterman, 1981). DP compares every pair of residues between two sequences and builds a matrix of substitution scores that represents all possible alignments. Matches and mismatches are scored using a substitution matrix and gaps are

penalized. The algorithm then produces the optimal alignment by tracing through the matrix and identifying and joining the highest scoring segments. DP has been proven to produce the mathematically optimal alignment (Lipman, et al., 1989), yet its execution requires computation and memory proportional to the square or cube of the sequence lengths. Despite the appeal of finding the optimal alignment, the slower performance of DP limits its application.

Significance

The development of statistical methods for estimating the significance of a pairwise alignment marked a major advance in protein sequence analysis. Determining and confirming the biological importance of a particular alignment is a difficult task, especially for proteins with low sequence similarity. Searches against sequence databases containing millions of sequences exacerbate this problem by producing numerous alignments. Thus, modern tools first estimate the significance of each alignment and filter out hits that fall below a user-specified threshold. The significance of a local alignment score may be determined given the distribution of alignment scores between completely random and unrelated sequences that have similar length and composition to the input sequences. Such random alignment scores follow a Gumbel extreme value distribution. If a particular alignment score is much greater than that expected by chance as determined by comparing to this extreme value distribution, the alignment is considered to be significant. The probability of an alignment occurring by chance, also known as the E-value, is determined using the Karlin-Altschul equation (Altschul, et al., 1994; Altschul and Gish, 1996; Altschul, et al., 1990). Low E-values indicate a significant alignment and *vice versa*. The statistical mathematics and complexities underlying the derivation of the Karlin-Altschul equation and other related equations are outside the scope of this dissertation.

Database searches

Popularized by its rapid searching method and ability to statistically rank pairwise alignments, the Basic Local Alignment Search Tool (BLAST) quickly became the *de facto* standard for performing sequence similarity searches. Unlike DP, BLAST does not find the optimal alignment, but rather approximates the optimal alignment using a heuristic search method. Specifically, BLAST locates all common *k*-mers, a contiguous subsequence of length k, between two sequences that score above a certain threshold. All the highest scoring segments are then extended and linked using DP and the resulting alignments estimated for their significance using Karlin-Altschul statistics. BLAST displays only those pairwise alignments scoring above a user-specified E-value. By restricting DP to sequence segments with promising similarity, BLAST bypasses searching the entire sequence space and thus overcomes the speed and memory limitations of DP. BLAST made possible the rapid querying of large sequence databases for homologous sequences (Altschul, et al., 1990). Consequently, the first clue to determining a novel protein's function often involves "BLASTing" its sequence against the SwissProt (Boeckmann, et al., 2003), non-redundant, and/or other protein databases. Furthermore, most automated annotation pipelines attempt to assign a particular function to each protein of a new proteome using BLAST (Fraser, et al., 2000).

Position Specific Iterative (PSI)-BLAST is a sensitive sequence similarity search tool that uses an iterative searching method and unique scoring scheme to detect weakly related homologs (Altschul, et al., 1997). The first iteration is equivalent to a normal BLAST search, and PSI-BLAST scores alignments using a standard substitution matrix (e.g. BLOSUM62). After each additional iteration, PSI-BLAST scores all pairwise alignments using a position-specific scoring matrix (PSSM) that was derived from a multiple alignment of the significant hits from the previous iteration (see below). The search process continues until convergence or when no new homologs are retrieved. PSI-BLAST is quite sensitive because the scoring matrix dynamically adapts to the observed amino acid substitution rate specific to each search and information from multiple

sequences is used rather than a single sequence. However, PSI-BLAST often requires human input in order to prevent unrelated hits from "corrupting" the scoring matrix and incorrectly biasing the search toward false positives. The ability to recognize weak sequence patterns makes PSI-BLAST a critical tool for comprehensively identifying all the members of a protein family or instances of a functional domain. Because the input and output domains of signal transduction proteins display low sequence similarity, their complete identification mandates the use of PSI-BLAST.

Multiple Sequence Alignment

Multiple sequence alignments provide an invaluable source of information into characterizing a group of related sequences. Alignments of protein sequences may be used to identify conserved amino acid patterns (Bork, et al., 1994), define functional and structural domains, identify novel members of a protein family (Eddy, 1998), build phylogenetic trees (Saitou and Nei, 1987), structure prediction (Rost, et al., 1994), and assist in experimental investigation. Despite their utility, constructing high-quality multiple alignments for divergent sequences (of less than 30% identity) is a laborious process often requiring the combination of computational tools and manual editing. Oftentimes in characterizing a functional domain, hundreds and sometimes thousands of sequences need to be aligned. In theory, producing an optimal multiple alignment is possible using a generalized version of the DP pairwise alignment algorithm; however, in practice, the space and time complexity of DP limits this approach to a small number of sequences (Lipman, et al., 1989). As a result, multiple alignment algorithms heuristically approximate the optimal alignment using either a progressive, consistency-based, iterative, or structure-based alignment strategy. Most tools adopt a hybrid approach that combines one or more of these alignment strategies.

Methods

Progressive alignment performs a pairwise alignment comparison of all sequences to each other using DP and then progressively aligns the most similar sequences or groups of sequences until no more sequences remain. The major disadvantage of this method is that the outcome strongly depends upon the order in which the sequences are aligned and propagates errors to any remaining sequences that have yet to be aligned. Even so, progressive alignment tends to execute rapidly, handle large numbers of sequences, and provide a relatively acceptable alignment that may serve as the starting point for further manual modification. The most popular tool for progressively aligning either DNA or protein sequences is the ClustalW program (Thompson, et al., 1994). Consistency-based measures use information from the pairwise alignments of all input sequences to determine the order of a progressive alignment and to appropriately weight the pairing of any two residues. T-Coffee (Notredame, et al., 2000) demonstrates this procedure and works well with less conserved sequences at the cost of a significantly longer execution time and memory footprint. Iterative strategies first create a draft alignment and then continue to refine this alignment until it surpasses a certain threshold. MUSCLE iteratively aligns large numbers of sequences in a reasonable time and the authors claim comparable results to T-Coffee (Edgar, 2004).

One of the goals of creating a multiple alignment for a particular protein domain is to align the residues with structural equivalence (i.e. each column will contain residues that support a particular structural feature such as an alpha helix or beta strand). Because a multiple alignment is expected to model the common structure of a group of related proteins, one recent method (Zhou and Zhou, 2005) for aligning proteins first predicts the secondary structure of each protein using PSIPRED (Jones, 1999) and then incorporates this structural information into the alignment process. This method demonstrated a 7-15% improvement in alignment accuracy for divergent (less than 30% identical) sequences.

Protein Domains

Multiple sequence alignments serve as the basis for representing and modeling protein domains. There are two types of protein domains: structural and homologous. A structural domain is the unit of a protein that can independently fold into a distinct, functional, 3D structure. Such structural units have identifiable patterns of amino acid conservation in sequence-space that are recognizable by computational protein sequence analysis.

Homologous protein domains are based on sequence characteristics from a group of closely related sequences that have been multiply aligned. Presumably, the multiple alignment reveals which amino acids are permitted in a given column and reflects the amino acid's structural function.

Sound statistical principles confirm that protein domains can be robustly represented by profile hidden Markov models (HMMs) (Eddy, 1998). A typical HMM is a linear series of match, insert, and delete states that are analogous to the similar characters and gaps of multiple sequence alignments. Associated with each state is a set of transition probabilities and a set of character probabilities. The former represents the likelihood of changing to a different state (e.g. one amino acid substituting or deleting for another), and the latter represents the probability of each character (in this case, amino acids) occurring in this state. These probabilities are calculated from a training set of homologous sequences that have been aligned using programs such as ClustalW (Thompson, et al., 1994) or MUSCLE (Edgar, 2004). After the training process, the HMM contains information regarding the conserved primary structure of the homologous domain sequences. In addition, each HMM may have user-supplied thresholds incorporated into its data structure which specify statistical cutoffs for delineating true members of a particular domain. Scanning a sequence database with each HMM may then identify novel domain homologs. Collections of such HMMs are stored in the primary domain databases, Pfam (Bateman, et al., 2004) and SMART (Letunic, et al., 2004).

Apart from identical sequences or experimental evidence, it is not ultimately possible to predict with certainty that two similar proteins will function the same. Therefore, while sequence similarity and comparisons form the core of comparative genomics, it is the integration and synergy of several independent lines of genomic (and experimental, when available) evidence that substantiates a functional prediction. Additional methods that were helpful in this research include the following: sequence similarity trees, sequence logos, and genome neighborhood analysis. Sequence similarity trees are useful for depicting the relationship between several sequences using a binary tree diagram. Because similar sequences cluster together, these trees visualize the distribution of related sequences and may reveal potential subfamilies with distinct functions (Saitou and Nei, 1987). Sequence logos graphically visualize the conservation of each position in a multiple alignment. This highlights conserved residues that may play an important functional role (Crooks, et al., 2004). Finally, adjacent genes are often co-expressed suggesting a common cellular role and signal transduction genes oftentimes are located beside the operons that they regulate. Therefore, the genomic neighbors for a particular gene or protein of interest may reveal significant clues as to a protein's function (Dandekar, et al., 1998; Overbeek, et al., 1999).

Comparative Genomics of Microbial Signal Transduction

The explosion of genomic data and steady increase of uncharacterized proteins necessitates the development of improved computational techniques for accurate prediction of protein function. Current automated annotation procedures often fail to adequately describe signal transduction proteins due to the high sequence variability of input and output domains and their mosaic domain architecture. Most signaling proteins in public databases are poorly annotated as "putative two-component sensor" or "two-component response regulator" partly because they are based on crude sequence similarity searches that reveal similarity to the conserved transmitter and receiver

domains. Furthermore, there is no comprehensive electronic resource or database for microbial signal transduction. The study of signal transduction could benefit greatly from advanced comparative genomic studies due to the numerous experimentally characterized systems and the lack of sufficient methods for automatically producing high quality annotations of signal transduction proteins.

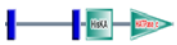
At this time, prokaryotes are the best model for using computational methods for the analysis of signal transduction systems. First, signal transduction pathways are much simpler in prokaryotes (typically, one to four protein components) than in eukaryotes (branched multi-protein cascades). Second, prokaryotic genes are often grouped together in operons, and signal transduction proteins are often associated with operons that they regulate. This makes genome context analysis a useful method for revealing the function of prokaryotic proteins. Third, more than 300 prokaryotic genomes are currently available (versus twenty-eight eukaryotic genomes) that span much longer evolutionary distances (which is critical for comparative analysis) than eukaryotes,. Fourth, prokaryotic gene-finding tools are considerably more sensitive and specific than eukaryotic gene-finding tools. This provides a much more comprehensive and accurate genetic record when working with multiple genomes. The smaller size of microbial genomes facilitates the management of genomic data. Finally, the well-established signal transduction community provides a source of experimental input – a critical link in ascertaining the validity and application of computational predictions.

Signaling Domains

The combination and type of signaling domains in a signal transduction protein determines its specific function. Therefore, reliable prediction of function for a signal transduction protein can only be derived from its complete domain architecture. The comparative genomics of microbial signal transduction largely comprises characterization of these signaling domains and interpretation of the cellular role of signal transduction

pathways at the domain level. Signaling domains belong to one of four classes of domains: input, transmitter, receiver, and output. They are principally modeled using HMMs (see above), which enables their rapid identification in protein sequences using the HMMER program. Preliminary analysis demonstrated that the majority of signal transduction proteins have partially characterized domain architectures with one or more regions where no domains are identified. The regions in prokaryotic protein sequences that might contain a domain are typically 80-100 amino acids long (Koonin and Galperin, 2003). Therefore, regions of eighty amino acids or longer that lack any statistically significant automated domain prediction, putatively contain an unknown or undetected domain. Examples of such undetected domains are shown in Table 1.2, which contains data for all six sensor histidine kinases encoded in the genome of the human pathogen *Campylobacter jejuni*. Each of these sensor proteins contains at least one undetected domain in their N-terminal input region, as revealed by PSI-BLAST searches against the non-redundant database. Nonetheless, these domains are not detected in automated searches against SMART (Letunic, et al., 2004) and Pfam (Bateman, et al., 2004). Therefore sensory function for any of these proteins was not predicted, and the proteins are currently incompletely and inconsistently annotated.

Table 1.2 Sensor histidine kinases from the genome of *Campylobacter jejuni*.

GI number	Domain architecture	Annotation (NCBI)
15792131		signal transduction histidine kinase
15792219		putative sensory transduction histidine kinase
15792546		putative two-component sensor
15792550		putative two-component sensor
15792586		two-component sensor (histidine kinase)
15792807		putative two-component sensor

Transmitter and Receivers

The homologous domains that comprise transmitter and receiver modules of signal transduction proteins are well conserved in sequence. Both SMART and Pfam contain high-quality HMMs for these domains. Thus, histidine kinases and response regulators are easily detectable by HMMER and related programs (Eddy, 1998). Scanning against Pfam and SMART has become a standard procedure for whole-genome annotation, and this has resulted in the successful identification of two-component signal transduction systems in completely sequenced microbial genomes (Grebe and Stock, 1999). Detecting transmitter and receiver modules in signaling proteins is an important and necessary step, but it does not provide answers to the most important biological questions: (1) What signal is detected by a given signal transduction system? and (2) What adaptive response does it produce? The detection of the input and output modules is necessary to provide the answers to these fundamental questions.

Input and Output Domains

Unlike transmitter and receiver domains, input and output modules are significantly less conserved in sequence. Because input domains must respond to such a broad range of signals, they are particularly variable and more difficult to identify by computational means. Relatively few input domains specific to signal transduction have been characterized, but they include the following: PAS (Ponting and Aravind, 1997; Zhulin and Taylor, 1997), GAF (Aravind and Ponting, 1997), Cache (Anantharaman and Aravind, 2000), CHASE (Anantharaman and Aravind, 2001; Mougel and Zhulin, 2001), CHASE2 through CHASE6 (Zhulin, et al., 2003) and NIT (Shu, et al., 2003).

Output domains are more highly conserved than input domains. The most commonly found output domains are DNA-binding helix-turn-helix (HTH) domains because the predominant adaptive response from response regulators is control of gene expression, which is mediated by binding to DNA. Several novel output domains have been recently described in response regulators, which implicate these systems in other types of control, such as the regulation of enzyme activity. These include adenylate and diguanylate cyclases, c-di-GMP-phosphodiesterase, phosphohydrolase, and other related domains (Galperin, et al., 2001; Nikolskaya and Galperin, 2002; Pei and Grishin, 2001; Shu and Zhulin, 2002).

Computational Characterization of Signaling Domains

All of the previously mentioned input and output domains were detected by using the Position-Specific-Iterative (PSI) BLAST program (Altschul, et al., 1997), a sensitive sequence-similarity search tool that requires manual control of parameters such as the inclusion threshold, low-complexity filter, composition-based statistics, and exhaustive searches seeded with any new homologs detected after each iteration (Altschul and Koonin, 1998; Koonin and Galperin, 2003; Schaffer, et al., 2001). These factors make PSI-BLAST a time-consuming and lengthy procedure. Furthermore, the follow-up work

describing a particular domain family involves construction of multiple alignments, identification of conserved residues and their contribution to a known or predicted structure, and determination of the domain architectures of related proteins. Thus, identification and characterization of novel domains in signal transduction proteins is a laborious procedure involving a variety of bioinformatics tools applied on a case-by-case basis.

Examples of Signaling Domains

PAS

The PAS domain is one of the most ubiquitous and important input domains in signal transduction (Taylor and Zhulin, 1999). PAS domains comprise a widespread domain superfamily, which are responsible for detecting light, oxygen, and redox potential. The computational characterization of PAS (Zhulin and Taylor, 1997) involved exhaustive PSI-BLAST searches using sequence data from the oxygen receptor, Aer, of *E. coli* (Rebbapragada, et al., 1997). Both Pfam and SMART contain models for the PAS domain and local HMM searches with this domain profile reveal more than 3500 signal transduction proteins containing this domain (October, 2005).

CHASE and NIT

Two other input domains recently described are the CHASE and NIT domains. Found within both prokaryotes and eukaryotes, the extracellular CHASE domain is predicted to bind small ligands (Mougel and Zhulin, 2001). This novel domain was characterized using exhaustive PSI-BLAST searches combined with multiple alignment construction, secondary structure prediction, identification of transmembrane regions in protein sequences, and analysis of amino acid conservation patterns. Essentially the same computational tools and methods were used for the identification of the NIT domain (Shu, et al., 2003). NIT containing proteins are predicted to carry out a very specific

function: detection of nitrate and nitrite in the extracellular environment. This prediction was made by extrapolating detailed biochemical data available for one of the homologs (Chai and Stewart, 1998).

ANTAR

Based on the analysis of an available three-dimensional structure, PSI-BLAST searches and other computational tools, Shu and Zhulin (2002) described ANTAR, an RNA-binding, output domain that is present exclusively in signal transduction proteins. The experimentally derived structure and function of this domain permitted the characterization of this domain despite only 5% sequence identity among sequence members. ANTAR carries out an important step in transcription anti-termination; therefore its identification in several human pathogens, such as *Mycobacterium tuberculosum*, *Listeria monocytogenesis*, *Burkholderia fungorum*, *Klebsiella pneumoniae*, provides important information that may be used in antimicrobial drug design.

Although in many instances the exact sensory or regulatory capabilities of a novel domain cannot be predicted directly, solving the complete domain architecture of proteins that contain this domain provides important clues about the biological function of the protein. Recently five novel input domains (CHASE2 through CHASE6) of unknown ligand-binding capabilities were identified (Zhulin, et al., 2003). By defining the complete domain architecture of related proteins containing these domains it was demonstrated that in bacteria similar signals can be transmitted via different signal transduction pathways to different regulatory circuits. The pervasiveness of a particular input domain in receptors from diverse regulatory networks in a given organism suggests the importance of a signal recognized by these receptors. These results provide evidence that refined biological function for signal transduction proteins can be predicted by solving the complete domain architectures.

Objectives

This research aims at reducing or closing the existing gap between genomic and experimental data in the area of signal transduction in the simplest organisms - prokaryotes. To achieve this goal, we pursued three major objectives. First, we built a customized and integrated bioinformatics platform to facilitate the genome-wide analysis of signaling systems. Second, using this platform, we sought to identify and analyze signal transduction proteins to investigate overarching trends throughout microbial signal transduction. Thirdly, we performed in-depth characterization of two sensory domains to provide a deeper understanding of specific signaling systems and to extrapolate this information onto existing and novel genomes.

CHAPTER 2

GENERAL MATERIALS AND METHODS

Since the inception of bioinformatics research, numerous tools and databases have been designed and they vary quite dramatically in complexity, scope, and application. Because bioinformatics research largely relies on comparisons, countless databases (Galperin, 2005) have been developed to manage widely diverse data sets. These range from raw collections of sequence data to highly specialized databases such as the EcoCyc database (Keseler, et al., 2005), which is specifically designed for managing information about the gram-negative bacterium, *E. coli*. In order to gain the maximal utility from these databases, many software programs have been designed to derive novel, additional information from primary data (such as a gene locus or protein sequence). The resulting output which itself might be incorporated into a secondary derivative database. The integrated use of multiple approaches requires insight and familiarity with the databases and their associated tools, so that they be efficiently applied to a particular biological problem. In this chapter, relevant bioinformatic tools and databases are described that aided in the research presented in this dissertation.

Databases

Three major types of databases discussed in this section:

1) Sequence databases that are simply repositories of either DNA or protein sequences with minimal annotation,
2) Structural databases containing biological macromolecular information, and
3) Domain databases that characterize various protein families.

Sequence Databases

GenBank

GenBank is a public database of nucleotide (DNA) sequences maintained by the National Center for Biotechnology Information (NCBI) as part of the National Institutes of Health. GenBank currently contains over 100 gigabases (1 x 10^11) of nucleotide data sequenced from more than 165,000 organisms (http://www.nlm.nih.gov/news/press_releases/dna_rna_100_gig.html, August 2005). GenBank continues to grow exponentially due to improved methods for sequencing new organisms and sequences from nearly 2000 new species are added every month. Traditionally, GenBank was partitioned based on taxonomy. Newer divisions such as the whole genome shotgun division, have been created to accommodate data from high-throughput sequencing initiatives. GenBank participates in the International Nucleotide Sequence Database Collaboration, which consists of GenBank, the European Molecular Biology Laboratory Data Library, and the DNA Data Bank of Japan. As part of this worldwide effort, these organizations exchange nucleotide data on a daily basis.

Each nucleotide sequence receives a stable identifier called an accession number and another unique identifier known as the GenBank Identifier (GI). While the accession number does not change, any updates to the sequence results in a new GI number and increments the record version. Incorporated into the annotation of each record are protein translations, features, and references. There is significant redundancy within GenBank owing to its comprehensive collection of DNA from multiple sources. GenBank is publicly available and may be accessed via the internet with services such as Entrez found at NCBI's site (http://www.ncbi.nlm.nih.gov/Entrez/) or downloaded from ftp://ftp.ncbi.nih.gov (Benson, et al., 2005).

NR – Non-Redundant Database

The NCBI non-redundant database (NR) is a comprehensive, non-redundant set of all protein sequences contained in GenPept, SwissProt, the Protein Information Resource, the Protein Data Bank, and the Reference Sequence database. In practice, NR contains some redundancy due to sequencing errors and minor differences between closely related proteins; however, it represents a good approximation at compiling identical sequences into a single entry. Currently, NR contains approximately three million proteins and represents the most complete "snapshot" of all available protein sequences. NR may be downloaded from NCBI at ftp://ftp.ncbi.nih.gov/blast/db/FASTA/nr.gz.

RefSeq – Reference Sequence Database

The NCBI Reference Sequence (RefSeq) database is a public, non-redundant database of genomic data, transcripts, and protein sequence information with the goal of providing an updated synthesis of both primary and secondary information for each sequence record. RefSeq is characterized by three features: 1) it is a curated, non-redundant collection of sequences; 2) it contains explicitly linked protein and nucleotide records to external sources of information; and 3) it represents significant taxonomic diversity with data from viruses, prokaryotes, and eukaryotes. RefSeq contains over one million protein sequences from more than 2,400 organisms. The annotation of each sequence is distinct from the original submission to GenBank with information gathered from multiple sources including: the original GenBank record, automated computational analyses, collaborators, manual curation, and user feedback. Similar to GenBank, each sequence is given a stable accession number and unique GI number. The RefSeq database is available for download at ftp://ftp.ncbi.nih.gov/refseq/release (Pruitt, et al., 2005).

SwissProt and TrEMBL

The SwissProt database is a high-quality collection of manually annotated, protein sequences and strives to provide state-of-the-art information for each sequence by combining experimental results and computational analysis. At a minimum, each record in SwissProt must contain the amino acid sequence, the protein name, its taxonomic classification, and a literature reference. Additionally, SwissProt contains extensive cross-references to external databases and each record follows a strict naming convention, which facilitates the searching and retrieval of relevant data. Identical sequences and their annotations are merged together to provide a low-level of redundancy. The SwissProt database places a special emphasis on protein sequences from *Homo sapiens* and model organisms with the goal of creating a representative collection of sequences; however, any data from specialized study groups, genomic-related publications or other literature highlighting a protein's function is also included in SwissProt. The current release of SwissProt (release 48.0, 13 Sep 2005) contains 194,317 records.

With the onset of improved sequencing technology and especially high-throughput genomic sequencing, many more protein sequences are being published than can be analyzed experimentally. Thus, the TrEMBL database includes all protein sequences (excluding those sequences already in SwissProt) from the CDS of nucleotide databases with a minimal level of annotation based solely on automated analyses. The current release of TrEMBL (release 31.0, 13 Sep 2005) contains 2,105,517 records (Boeckmann, et al., 2003). Information on downloading SwissProt and TrEMBL may be found at http://us.expasy.org/sprot/download.html.

Structural Databases

PDB – Protein Data Bank

The Protein Data Bank (PDB) is the primary, centralized resource for three-dimensional structures of biological macromolecules. Each structure included in the PDB contains the organism name, SwissProt and GenBank identifiers, PubMed identifiers, and enzyme commission numbers. From this information, links are automatically generated to taxonomy information at NCBI, gene ontology terms, structural genomic targets, and when relevant, to the Kyoto Encyclopedia of Genes and Genomes enzyme database. The PDB stresses data integration and linking to other relevant databases. The PDB may be downloaded from ftp://ftp.rcsb.org/pub/pdb/ in a variety of formats (Deshpande, et al., 2005).

Domain Databases

Pfam – Protein Family Database

Pfam is a database of protein families and domains. Each domain is represented with a multiple alignment and profile hidden Markov model (HMM), and annotated with a text description, references to pertinent literature, and links to external resources. All entries are classified into one of four categories:

- *Family*: a related group of protein sequences,
- *Domain*: a structural unit able to exist independently,
- *Repeat*: a unit that only occurs in two or more copies, and
- *Motif*: a shorter sequence unit not found as part of a globular domain.

Pfam is comprised of Pfam-A, a manually curated set of domains with detailed information about each family, and Pfam-B, an automatically generated set of domains based on motifs from the PRODOM database (Bru, et al., 2005). The Pfam database is freely available in a variety of formats including HMM's for local searching with the HMMER software (see below), multiple alignments, and the annotation of each family. Because of its high-quality domain families and easy extrapolation of this information onto any given protein sequence, Pfam is an invaluable resource in large-scale functional annotations. In August 2005, the Pfam curators released version 18.0 which contained 7973 families.

The Pfam web server (http://www.sanger.ac.uk/Software/Pfam/) provides a number of services for analyzing a protein sequence over the Internet. This includes predicting Pfam domains, signal peptides, transmembrane regions, coiled-coil segments, and low-complexity regions. Complex search options are available such as scanning by taxonomy and looking for particular domain organizations (Bateman, et al., 2004).

SMART – Simple Modular Architecture Research Tool

SMART (Simple Modular Architecture Research Tool) is a database of signaling domains and focuses on characterizing eukaryotic proteins, although numerous prokaryotic domains have also been characterized. Similar to Pfam, each domain is modeled with an HMM and multiple sequence alignment along with descriptive information including publications, functional assignments, and distribution across phyla. Upon signing a licensing agreement, the SMART models may be downloaded for local searching using the HMMER software (see below). The SMART website (http://smart.embl-heidelberg.de/) provides enables the user to scan protein sequences for

both Pfam and SMART domains and will identify other sequence features such as transmembrane regions. The January 2004 release of SMART contained 685 domain families (Letunic, et al., 2004).

COG – Clusters of Orthologous Groups

The COG database attempts to classify all conserved genes from available complete genomes into clusters of orthologous groups or COGs. Three or more proteins from distantly related lineages are considered orthologs if they are more similar to each other than any other proteins from the same genome. Because orthologs are considered to share the same function, other uncharacterized or hypothetical genes assigned to a COG may be considered to have an identical function. The COG database has been successfully used to annotate new genomes, characterize a vast number of hypothetical conserved genes, and identify phyletic patterns between various taxonomic groupings (Tatusov, et al., 1997).

Tools

Sequence Similarity Search Tools

BLAST – Basic Local Alignment Search Tool

The first evidence for homologous sequences to a given protein or nucleotide sequence usually begins by searching for similar sequences in a sequence database using BLAST or one of its variants. BLAST rapidly and efficiently identifies statistically significant, pairwise alignments of a query sequence to any other sequences in a target database. The BLAST algorithm accomplishes this feat by approximating the optimal alignment rather than performing a time consuming search of the entire sequence space.

This avoids the construction of many insignificant alignments. While this may result in the loss of some sensitivity, the computational runtime is improved by at least a magnitude of order. Thus BLAST trades some sensitivity in order to gain a huge reduction in the computational time.

The BLAST algorithm comprises three major steps: seeding, extension, and evaluation. Seeding the search space involves detecting all the word-hits, or seeds, between two sequences. A word is a contiguous subsequence of some length W. A word-hit occurs when two sequences share a word that scores at least T as defined in a substitution matrix such as BLOSUM62 (Henikoff and Henikoff, 1992) or PAM250 (Dayhoff, et al., 1978). The seeding step is followed by extending each word-hit in both directions until the alignment score falls X units below the maximum score of the alignment generated thus far. The original version of BLAST performed ungapped alignments; however, the latest version computes gapped alignments in the extension step using dynamic programming (Altschul, et al., 1997). Finally, BLAST evaluates which of the alignments generated by extending the seeds are statistically significant. After sorting the alignments by score, the significance of each alignment in terms of its E-value is determined by the Karlin-Altschul equation. Though the statistical mathematics and complexities underlying the Karlin-Altschul equation and related equations are outside the scope of this dissertation, the E-value represents the likelihood that an alignment would occur randomly. It is calculated from the score, size of sequence search space, and a minor constant. A lower E-value corresponds to a more significant alignment. Any alignments that score above a user-specified E-value are termed high-scoring segment

pairs and are presented to the user along with other relevant information including the database search parameters (Altschul, et al., 1990).

PSI-BLAST – Position Specific Iterative BLAST

PSI-BLAST (Position specific iterative BLAST) enables the detection of weakly related homologs via an iterative searching procedure that scores each alignment with a position specific scoring matrix (PSSM). A PSSM is a two-dimensional frequency matrix that captures the distribution of sequence characters found at each position in a multiple alignment. In other words, the PSSM characterizes the likelihood of finding a particular character at a specific position within a group of aligned sequences. The first PSI-BLAST round executes identically to BLAST using one of the supplied substitution matrices (e.g. BLOSUM62) to score the pairwise alignments. Each subsequent round of PSI-BLAST is preceded by the construction of a multiple alignment from any significant hits, and a corresponding PSSM is used in place of the query sequence to search and estimate the significance of any new hits. Searching continues until convergence (i.e. no more homologs identified) or some arbitrary threshold (e.g. maximum number of iterations) is exceeded. Because new sequences are evaluated based not on a standard scoring matrix but from sequences comprising the PSSM, caution must be taken when deciding which sequences to include for the next iteration in order to prevent "corrupting" the PSSM and thereby scoring unrelated or random sequences as significant (Altschul, et al., 1997).

BLAST databases

Before a sequence can be compared to other sequences using BLAST or PSI-BLAST, the target sequences to be searched against must be compiled into a BLAST database. NCBI regularly releases several, pre-formatted BLAST databases from

sequence databases such as RefSeq and NR. These may be searched via the NCBI BLAST server (http://www.ncbi.nlm.nih.gov/BLAST/) or downloaded for local searches using the BLAST executables. Custom BLAST databases may be created from a FASTA-formatted (Pearson and Lipman, 1988) file using the program formatdb. Both formatdb and the BLAST programs are part of the freely available NCBI-Toolkit (ftp://ftp.ncbi.nih.gov/toolbox/ncbi_tools/CURRENT/).

Multiple Sequence Alignment and Phylogenetics

The simultaneous alignment of multiple protein (or DNA) sequences reveals how a group of proteins are related in terms of sequence similarity. As such, multiple sequence alignments represent an invaluable source of information for many bioinformatic tasks including characterizing protein families, identifying functionally conserved residues, searching for remote homologs, and performing phylogenetic analyses. In this subsection, programs used to generate multiple sequence alignments and phylogenetic trees using various methodologies are presented.

ClustalW

ClustalW is a robust, multiple sequence alignment tool capable of rapidly aligning many nucleotide or protein sequences using a progressive alignment strategy. Initially, all sequences are aligned to each other to produce a distance matrix, which contains the distance from a given sequence to any other sequence in the group based on sequence identity. From this distance matrix, the ClustalW program builds a guide tree, which directs the order in which sequences should be sequentially aligned. Following the guide tree, ClustalW progressively aligns the sequences to each other beginning with the most related sequences and ending with the most divergent sequences. ClustalW has numerous

features and options, such as sequence weighting and adjusting gap penalties. Furthermore, ClustalW offers the capability to align two distinct multiple alignments to each other (Thompson, et al., 1994).

As with most multiple alignment programs, ClustalW performs well when aligning closely related sequences (greater than 30% identical or with regions of high similarity) and provides mediocre results for more distantly related sequences. The progressive strategy of aligning the most similar sequences first works well in many cases; however, alignment errors at earlier stages of the process are propagated downstream as the remaining sequences are analyzed. Despite these drawbacks, ClustalW executes very rapidly and can align large numbers (thousands) of sequences in a reasonable length of time even on commodity hardware.

T-Coffee

T-Coffee (Tree-based Consistency objective function for alignment evaluation) uses a progressive alignment strategy that takes into account both local and global alignments and information between all pairwise alignments to produce a multiple sequence alignment. The T-Coffee algorithm follows three major stages: 1) constructing a primary library from global and local alignments, 2) extending the primary library, and 3) progressively aligning the sequences using information from these libraries. Initially, global (performed by ClustalW (Thompson, et al., 1994)) and local alignments (generated by LALIGN (Pearson and Lipman, 1988)) are combined into a primary library with each pairwise alignment weighted by its sequence identity. Extending the primary library consists of further weighting each pair of residues by the frequency this pairing occurred in all the other pairwise alignments. Thus, the library contains information about pairwise

alignments between all input pairs rather than information localized to pairing the two most similar sequences. Finally, T-Coffee progressively aligns all the sequences based upon the position-specific weighting scheme of each residue pair from the extended library (Notredame, et al., 2000).

T-Coffee tends to align divergent sequences (less than 30% identical) more accurately than ClustalW. By pairing residues with information from all pairwise alignments, gaps also are more likely to be properly placed. Unfortunately, T-Coffee is heavily time and memory intensive, and it requires substantial hardware in order to align large numbers (more than 200) of sequences in a reasonable period of time.

PCMA

PCMA (profile consistency multiple sequence alignment) multiply aligns sequences using a progressive alignment strategy that balances speed and accuracy. PCMA first aligns all sequences with pairwise identity greater than some threshold (e.g. 40%) using ClustalW. The remaining sequences and groups with lower sequence similarity are aligned using T-Coffee. The authors of PCMA reported accuracy similar to T-Coffee, yet exexcution time is twenty times faster.

Most multiple alignment tools produce relatively accurate alignments of closely related sequences (i.e. greater than 35% sequence identity), but this accuracy is substantially reduced when attempting to align more divergent sequences. Furthermore, techniques such as the T-Coffee algorithm, which perform reasonably well with divergent sequences, tend to require large amounts of time and memory. By combining the ClustalW and T-Coffee algorithms into a hybrid approach for aligning sequences, PCMA

provides an interface for rapidly aligning sequences without sacrificing the accuracy of T-Coffee (Pei, et al., 2003).

MUSCLE

MUSCLE builds multiple sequence alignments through *k*-mer distance estimation, progressive alignment using a novel log-odds expectation score, and a tree-dependent restricted partitioning refinement process. In brief, MUSCLE builds a draft progressive alignment based upon a distance matrix created using a *k*-mer distance measure (Edgar, 2004). The draft alignment is then improved by first recalculating the distance matrix with the Kimura distance measure (Kimura, 1985) and progressively aligning the sequences based on this new model. Finally, MUSCLE refines the alignment using tree-dependent restricted partitioning – an iterative improvement process which divides the alignment based on splitting the tree into two subtrees at a random position and realigning the resulting profiles. This continues until convergence or some user-specified threshold is exceeded. MUSCLE claims to be the fastest and most accurate of ClustalW, T-Coffee, and MAFFT (Katoh, et al., 2002) when compared with four, reference alignment test sets (Edgar, 2004). As such, MUSCLE represents one of the best multiple alignment tools available.

SPEM

SPEM constructs protein multiple sequence alignments using profile-profile alignments and predicted secondary structures. For each protein sequence, the secondary structure of each protein is predicted using PSIPRED (Jones, 1999) and a sequence profile derived using PSI-BLAST. SPEM then generates all pairwise alignments using DP, secondary-structure dependent gap penalties, and secondary structure profile

information. From these pairwise alignments a guide tree is built using the neighbor-joining method (Saitou and Nei, 1987) to progressively align all the protein sequences relative to a position specific scoring scheme. SPEM reported an average 7-15% higher accuracy than T-Coffee, MUSCLE, and Probcons (Do, et al., 2005). Due to the amount of time involved in building a PSI-BLAST profile and predicting the secondary structure for each protein, SPEM is requires more computational resources than T-Coffee. Thus, SPEM provides relatively high accuracy and is suitable for aligning medium size numbers of divergent sequences (Zhou and Zhou, 2005).

MEGA

MEGA (molecular evolutionary genetics analysis) is a software package to facilitate the investigation of DNA and protein sequence information in an evolutionary context. The latest release, version 3.0, contains numerous capabilities to analyze molecular sequences both manually and automatically including the following: sequence acquisition, alignment, estimating evolutionary distances, and building phylogenetic trees (Kumar, et al., 2004). MEGA was primarily used in this research to produce sequence similarity trees using the neighbor-joining method (Saitou and Nei, 1987).

Consensus

Consensus is a Perl based program for determining the consensus of each position within a protein multiple sequence alignment. Consensus tallies which residues of a column belong to various amino acid groupings (e.g. tiny – alanine, glycine, serine; negative – aspartic acid, glutamic acid; etc.) and displays a character representing that grouping if its representation exceeds a given threshold (e.g. 75%). Such an analysis can reveal functionally conserved sites by identifying various positions whose amino acid

composition suggests a functional role. Consensus is available on the Internet (http://www.bork.embl-heidelberg.de/Alignment/consensus.html) or as a Perl script for local execution.

Alignment Shader

Alignment Shader visualizes the conservation within user-specified subgroups of a protein multiple sequence alignment and enables the rapid assessment of conserved residues specific to a particular subgroup. If the BLOSUM consensus for each column of each subgroup exceeds a given threshold (e.g. 75%), Alignment Shader colors the background of the amino acids that belong to the BLOSUM consensus. Alignment Shader is incorporated as part of the bioinformatics platform (Chapter 3) and may be accessed at http://moscow.biology.gatech.edu/cgi-bin/alignshade.cgi.

Jalview

Jalview is a Java multiple alignment editor designed for the fast and efficient viewing and editing of large multiple sequence alignments. Jalview has numerous features and capabilities that make it a robust alignment tool including the following: data input and output in a variety of formats, building phylogenetic trees, coloring alignments using an informative color scheme, defining sequence groups, etc. Jalview also supports multiple distinct, integrated views of a sequence alignment (Clamp, et al., 2004).

Secondary Structure Prediction

PSIPRED

PSIPRED predicts the secondary structure of a protein sequence using neural networks and PSI-BLAST profiles. Initially, PSIPRED performs a PSI-BLAST search

with the query protein sequence against a filtered (to remove repetitive regions that would bias the PSI-BLAST search) NR database. PSIPRED then extracts the PSSM from the PSI-BLAST search and feeds this information into a neural network which assigns each residue of the query sequence to one of three secondary structure states: helix, strand, or loop – also known as the Q3 designation (Rost, et al., 1994). At the time of its publication, PSIPRED ranked as a leading secondary structure prediction tool based on its average Q3 score of 76.5 – 78.3% at the 1997 Critical Assessment of methods for protein Structure Prediction experiment (Moult, et al., 1997). Furthermore, PSIPRED remains a leading secondary structure prediction tool based as demonstrated by the evaluation of secondary structure prediction servers (Koh, et al., 2003).

The major principle behind PSIPRED is that sequence variability and conservation information from a group of related proteins enables a much more nearly accurate prediction than just the information from a single protein sequence. Regions of a protein sequence that have a specific function such as an active site or residues for maintaining the structural fold tend to display a high degree of conservation. More variable regions tend to exist on protein surfaces where fewer constraints exist except for a propensity for hydrophilic residues due to the surrounding polar environment. By collecting position specific, amino acid conservation information about related proteins using PSI-BLAST, PSIPRED assesses the solvent accessibility of each amino acid and predicts the secondary structure state based upon the solvent accessibilities typical for secondary structures (Jones, 1999).

Phobius

Phobius may be used to simultaneously predict signal peptides, transmembrane regions, and their topology from a protein sequence. Numerous other programs exist that separately predict transmembrane regions or signal peptides; however, since the patterns for these two secondary elements are highly similar they are prone to false classifications. Phobius' major strength lies in its ability to distinguish between transmembrane regions and signal peptides and to predict both elements in a single step. Furthermore, Phobius is taxonomy-agnostic, meaning that it was trained without respect to a specific kingdom. This enables the blind searching for signal peptides or transmembrane regions without having to know the taxonomy of a given protein sequence *a priori.*

On average, Phobius predicts signal peptides and transmembrane regions with the following accuracies: proteins containing both signal peptides and transmembrane regions, 90%; proteins with transmembrane regions only, 60%; proteins with signal peptides only, 96%; and proteins with neither signal peptides or transmembrane regions, 98%. (Kall, et al., 2004). When compared to SignalP (Bendtsen, et al., 2004) and TMHMM (Krogh, et al., 2001), Phobius predicted significantly fewer misclassifications although it was also slightly less sensitive. Phobius may be freely installed on Unix-based machines to academic institutions, and queries may be submitted online at http://phobius.cgb.ki.se/.

Coils

The Coils program identifies coiled-coil regions from a protein sequence. Coiled-coil regions typically consist of two to five helical bundles that twist around each other to form a supercoil and display a characteristic 'knobs-in-holes' packing of the peptide sidechains at the helix-helix interface. Proteins with coiled-coils demonstrate a heptad

repeat where the first and fourth residues (positions a and d) of every seven amino acids are hydrophobic while the other five residues (positions b,c,e,f, and g) are hydrophilic. Heptad repeats readily enable the prediction of coiled-coil regions and are often found in long filamentous proteins such as myosin or chemoreceptors (Lupas, 1997). The Coils program may be downloaded from ftp://ftp.ebi.ac.uk/pub/software/unix/coils-2.2/.

Seg

Seg filters out repetitive or regions with low-complexity from protein sequences. Many proteins contain short repeats, homopolymers, or significant subsequences consisting of just a few amino acids types. These uninformative segments can be quite problematic for many bioinformatic analyses such as sequence similarity searches (Wootton and Federhen, 1993). Seg is freely available from ftp://ftp.ncbi.nih.gov/pub/seg/.

Domain Architecture Prediction

HMMER

HMMER is a software package for manipulating and working with profile HMMs to analyze protein sequences. Profile HMMs are statistical models that have been successfully used to model protein domains (see introduction). The HMMER software package contains several tools for handling HMM-related tasks including the following: building HMMs from a multiple sequence alignment (hmmbuild), calibrating an HMM (hmmcalibrate), constructing domain libraries, searching sequence databases for domain homologs (hmmsearch, hmmpfam), and aligning sequences to an HMM (hmmalign) (Eddy, 1998). HMMER is an invaluable tool for determining the domain architecture of a

given protein. It has been implemented as a search tool at both the Pfam and SMART web servers and may be installed locally on Unix-based architectures for use with any HMM compatible database (e.g. Pfam, SMART, or customized HMM collections). HMMER is available at http://hmmer.wustl.edu.

Visualization

PyMOL, VMD

PyMOL and VMD (Visual Molecular Dynamics) are freely distributed, visualization programs for manipulating, analyzing, and visualizing three-dimensional biological structures. They both support a broad range of functions including multiple simultaneous views of one or more structures, powerful drawing routines of structural features at multiple levels of detail, scripting, measuring distances, and animations. Structures may be loaded from a variety of formats including the popular PDB format from the local hard disk or automatically from the online PDB database. More information about PyMOL including installation instructions is available from the PyMOL website, http://pymol.sourceforge.net/. VMD may be downloaded from http://www.ks.uiuc.edu/Research/vmd/ and is described in Humphrey, et al., 1996.

Weblogo

Weblogo creates sequence logos from a multiple sequence alignment. Sequence logos graphically represent the amino acid conservation at each position within a multiple alignment. The column height indicates the total conservation at this position and the height of each letter within the column denotes that amino acid's conservation (Crooks, et al., 2004).

Blender

Blender is a powerful 3D animation package that was used for creating complex 3D bar graphs and movies (http://www.blender.org).

CHAPTER 3

BIOINFORMATICS PLATFORM AND THE MIST DATABASE

Introduction

A thorough, computationally based analysis of microbial signal transduction necessitates the construction of a flexible, integrated, and unified platform for executing programs, storing and retrieving data, and investigating results. Currently, the bioinformatics field is saturated with numerous methodologies, data formats, heterogeneous databases, and incongruent tools. Such fragmentation impedes progress in comparative genomics where integration and comparisons are essential. To address this need for integrated and uniform data management, we deployed a complex assembly of bioinformatic tools, relational database, and knowledge environment. This was implemented using high quality hardware including a powerful dual-Xeon server and a 16-processor Linux cluster. Such an approach requires substantial effort to initially configure, yet it yields a consistent and reliable system that is insulated from external changes. In this chapter, we describe the hardware and software underlying our bioinformatics platform, the central Microbial Signal Transduction (MiST) database, the high-throughput process for identifying signal transduction systems, and the exploratory interface to this platform and MiST.

Hardware and Software

The bioinformatics platform was implemented on state-of-the-art hardware and technology, including a primary server and a 16-processor Linux cluster. The primary

server consisted of dual Intel-based 3.0 gigahertz (Ghz) Xeon processors, four gigabytes (GB) of random access memory (RAM), four SCSI hard drives with 584 GB of storage space that was configured in a redundant array of independent disks (RAID level ten for optimal performance and stability) and the Gentoo Linux operating system (OS). Gentoo (http://www.gentoo.org) is a source-based distribution of the Linux OS that provides maximal performance and customization. Connected to this primary server is an 8-node, 16-processor Linux cluster with a peak performance of approximately twenty GigaFLOPs (Floating-point Operations) and 120 GB of NFS-mirrored, storage space. Each node consists of dual Intel-based 2.2 Ghz Xeon processors, two GB of RAM, 120 GB of local storage, and the RedHat Linux OS (http://www.redhat.com).

In addition to the numerous open-source packages that come as part of the Gentoo and RedHat Linux distributions, the following software packages also comprised the bioinformatics platform:

- Relational Database Management System (RDBMS): MySQL (http://www.mysql.com), PostgreSQL (http://www.postgresql.org)
- Programming languages: Perl, Python, C/C++
- Support libraries: eXtensible Markup Language (XML), Common Gateway Interface (CGI), Library for the World-wide-web for Perl (LWP), Database Interface (DBI), Twisted (http://www.twistedmatrix.com)
- Web server: Apache 2.0 (http://www.apache.org)
- Cluster management: ROCKS (http://www.rocksclusters.org/Rocks/), Sun Grid Engine (SGE) (http://gridengine.sunsource.net)

The MySQL and PostgreSQL RDBMS were chosen for their substantial implementation of the Structured Query Language (SQL) specification, ease of deployment and integration with other tools (e.g. the Perl language, phpMyAdmin, and phpPgAdmin), functionally rich feature sets, high performance, portability, minimal overhead and maintenance, and stability. One of the most popular RDBMS for biological databases, MySQL, is known for its incredible querying speed, web integration, and straightforward configuration; however, due to MySQL's emphasis on speed, it lacks many features associated with other database systems. PostgreSQL constitutes a somewhat slower but more feature-rich RDBMS with support for foreign key references, triggers, views, and nested subqueries. Thus, MySQL is better suited for web applications where speed is critical, and PostgreSQL is preferred for structuring and querying complex data sets.

Perl is the preferred language for bioinformatics because of its scripting capabilities and superior text parsing. Perl also contains useful add-in modules for simplifying the tasks of database management (DBI) and web interfacing (CGI, LWP). For example, the DBI module reduces the complexity of interacting with MySQL or PostgreSQL to a few lines of code and the LWP module enables one to interact with websites from a single command. The Python programming language combines both the powerful simplicity of an interpreted language with the object-oriented capabilities often found in lower-level languages such as C++ or Java. Python in conjunction with the Twisted network library was primarily used as an intermediate networking layer for handling communications between the server and cluster. C/C++ were used to improve performance when the speed of the Perl and Python scripting languages became a

limiting factor; however, C/C++ are quite complex and require careful programming to avoid programming errors such as memory leaks and buffer overflows.

The Apache software package was used to deliver HTML (HyperText Markup Language) content and served as a portal for users to interact with the bioinformatics platform from the Internet. Additionally, customized CGI scripts were developed to execute computationally demanding tasks such as PSI-BLAST and HMMER on the Linux cluster. Scheduling, distributing, and execution of job requests on the cluster were handled using SGE. The cluster is configured and maintained with the Rocks Cluster software.

Whenever possible, data exchange between various programs and scripts was encoded in the XML format to facilitate parsing and to standardize input and output operations. This was particularly important when importing data originating from heterogeneous, external sources in various formats. Non-XML data must first be parsed and then formatted into XML adhering to a specific Document Type Definition (DTD). The DTD describes the components, structure, and organization of a particular set of data in a standard format by associating various XML tags with specific biological information.

Building onto this core infrastructure, several additional bioinformatic databases and tools were installed. These included: NR, RefSeq, SwissProt, Pfam, and SMART, BLAST, PSI-BLAST, ClustalW, T-Coffee, PCMA, MUSCLE, Consensus, PSIPRED, Phobius, Coils, Seg, HMMER, PyMOL, VMD, Weblogo, and Blender. Detailed information regarding these databases and tools may be found in chapter 2, “General Materials and Methods.”

In addition to the open-source software and common bioinformatics programs described above, we developed several utility scripts for performing tedious but useful bioinformatic tasks:

- aln2fasta – converts a ClustalW alignment into FASTA format
- alncdel – removes gaps from a ClustalW alignment
- alnidred – replaces the identifier lines of a ClustalW alignment with unique identifiers that are associated with the original identifier
- alnlong2short – transforms a ClustalW alignment from a single block of sequences into multiple blocks of sequences
- alnshort2long – transforms a ClustalW alignment from multiple blocks of sequences into a single block of sequences
- alnstripgaps – removes any columns that contain only gaps from a ClustalW alignment
- buildRefSeqFasta – downloads the complete RefSeq database in FASTA format
- crc – calculates the 64-bit cyclic redundancy check hash of FASTA formatted sequences
- fa2smart – automatically submits FASTA formatted sequences to the SMART website in order to predict their domain architecture
- faadups – removes any duplicate sequences from a FASTA formatted file
- facount – counts the number of sequences in a FASTA formatted file
- faidexp – replaces the unique identifiers of sequences in a FASTA formatted file with their original identifiers

- faidred - replaces the identifier lines of sequences in a FASTA formatted file with unique identifiers that are associated with the original identifier
- fasplit – splits a FASTA formatted file into either a user-specified number of files or number of sequences per file
- fixnr – concatenates identical sequences from the NR into a single entry
- gilst2fasta – downloads from NCBI the FASTA formatted sequence(s) for a list of GI numbers
- gilst2gbk – downloads from NCBI the GenBank record(s) for a list of GI numbers
- gbcount – counts the number of sequences in a GenBank formatted file
- hmmcount – counts the number of HMMs in an HMM database
- pmatch – scans a given file for a Perl-compatible regular expression and print any matches
- taxfromgi – prints the taxonomy for a list of GI numbers

The MiST database

The vast quantity of genomic data produced by sequencing and annotation projects necessitates the development of databases to organize, comprehend, and analyze this information effectively (Karp, 2001). Thus in order to perform a comprehensive comparative genomic analysis of microbial signal transduction, we designed and implemented the Microbial Signal Transduction (MiST) database using the PostgreSQL (version 8.0) RDBMS. MiST functions as the backbone of the bioinformatic platform by serving as a central, integrated repository of both primary and derived information. The structure of MiST is conceptually divided into four extensible sections (Figure 3.1):

- *Primary genomic data*: DNA and protein sequences and annotation information

- *Derived data*: predicted domain architectures and other secondary features (e.g. transmembrane and low-complexity regions) for each protein, and domain model information
- *Results*: signal transduction domains, classified signal transduction proteins, and statistical information
- *Management*: bookkeeping and tracking of genomic data mining

The primary genomic data section is composed of the Datasource, Organisms, Components, Proteins, and ProteinsData tables. Records within the Components table represent chromosomes, plasmids, or sequence contigs (for draft genomes). Associated with each component are related genes and proteins. Derived data is meant to store and organize any information derived from primary data by computational methods. Each protein's predicted domain architecture, transmembrane regions, signal peptides, coiled-coil regions, and low-complexity regions are stored in the Regions table.

Any proteins determined to be involved in signal transduction are stored in the Results section, which is comprised of the SignalProteins and SignalFamilies tables. Finally, meta-data regarding the status of information flow for each component is maintained in the Tasks table of the management section. The Tasks table identifies any component that has been downloaded to the local file system, parsed and inserted into the database, and tracks whether the domain architecture and various other secondary features have been predicted for the proteins related to each component. These database values are updated with the execution and completion of each task. This infrastructure facilitates the management of both storage and external processing of this data without restricting future development. Incorporating additional tasks simply involves creating

the supporting tables, relating the new tables to existing tables, and adding another column to the tasks table. Perl modules have been developed which interact with the database to enable the rapid construction of custom data files for exporting purposes or processing by other programs.

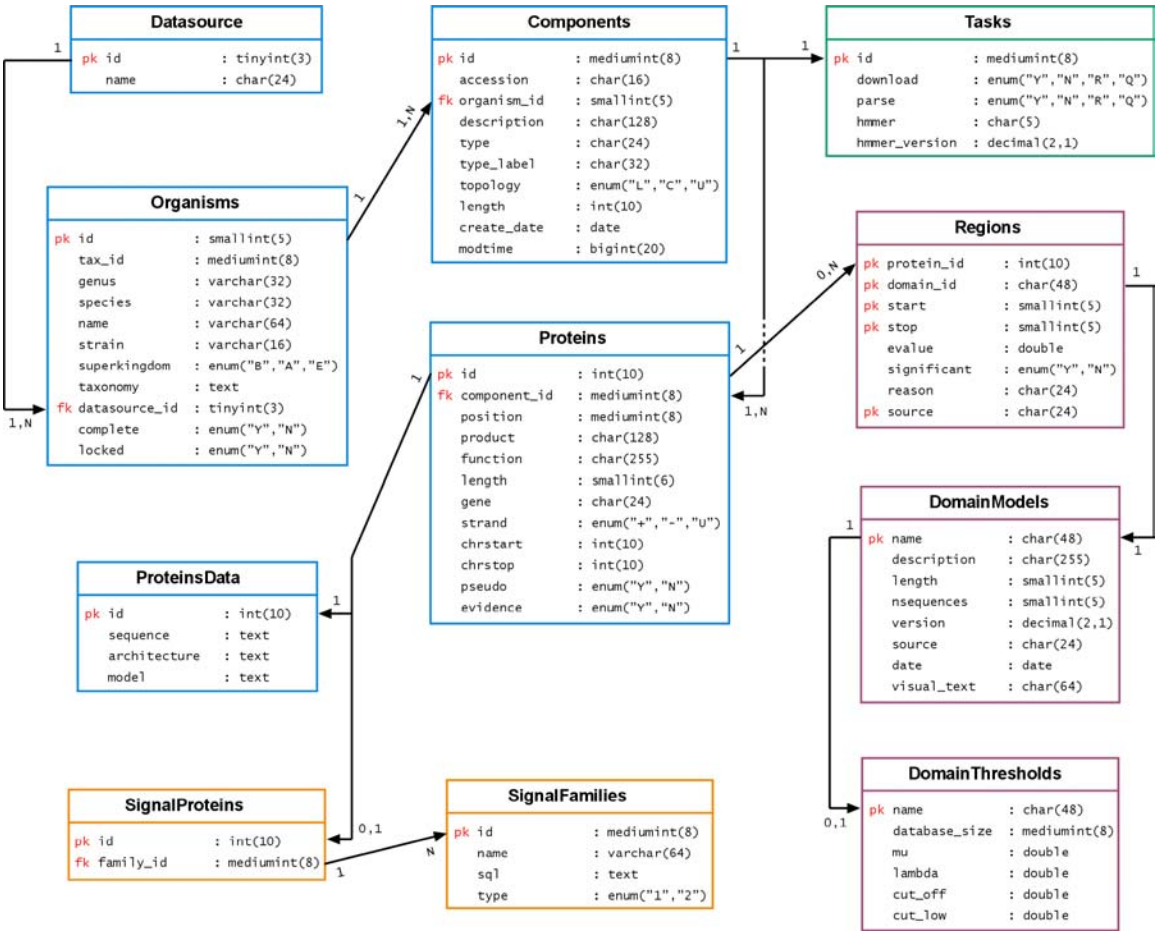


Figure 3.1 Structure of the Microbial Signal Transduction database, MiST. Table names are in bold type and lines indicate the various relationships between tables. Values above the lines signify the cardinality constraints for each relationship. The labels 'pk' and 'fk' refer to primary key and foreign key, respectively. In some cases, a single column is both a primary and foreign key, in which case only 'pk' is listed. The color of each table denotes a particular database section: light blue – primary genomic data; purple – derived data; orange – results; and green – management.

High-throughput identification of signal transduction proteins

The high-throughput identification of signal transduction proteins in prokaryotes involves three major steps (Figure 3.2):

1. All available genomic data is downloaded to the local file system, parsed, and inserted into MiST.
2. The domain architecture and various secondary features of each protein is computationally predicted and inserted into the database.
3. Finally, the set of signal transduction proteins is determined by querying each protein in the database for the presence of conserved signaling domains.

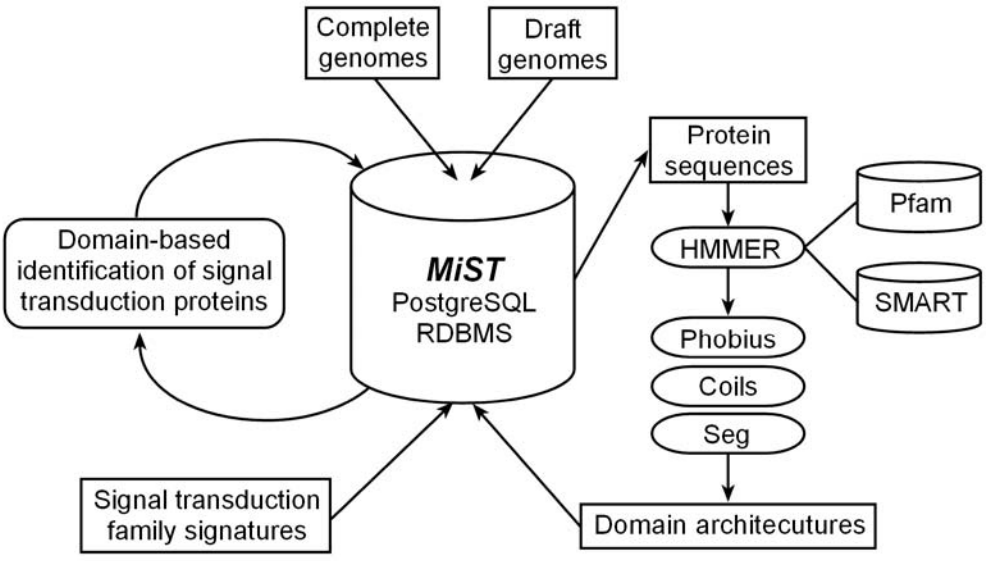


Figure 3.2 High-throughput identification of signal transduction proteins.

Step 1: Retrieval of genomic data and incorporation into MiST

All available complete and draft genomes are downloaded and processed locally so that subsequent information processing is minimally dependent on external sources. Draft genomes consist of several discontinuous shotgun sequences, which may contain overlaps and other sequencing errors. Despite the preliminary nature of this data, many experimental scientists are interested in receiving information about signal transduction proteins from these genomes. Moreover, preliminary analysis shows the largest and most diverse groups of signal transduction proteins are found in several of these draft microbial genomes (e.g. *Magnetospirillum magnetotacticum*). For that reason we have developed a computational infrastructure capable of handling both complete and incomplete genomic data.

Perl scripts take a "snapshot" of all available genomes from the NCBI web and ftp (file transfer protocol) sites, then store this information in XML for further processing. For each genome not present in MiST, its XML encoded genomic data is downloaded to a standard location on the local file system and meta-data regarding this genome's status recorded in the database. In addition to the nucleotide data and translated protein set, these XML files contain the full RefSeq annotation associated with each gene and protein. Using the XML::Parser module, Perl programs parse this XML data and insert this information into the appropriate tables within MiST.

Because Pfam is openly available and freely provides annotations with the domain models, this library was integrated by recording the model description and parameters in the Pfam tables of MiST. Similarly, any novel domain models may be added using this same process.

Step 2: Prediction and storage of each protein's domain architecture and secondary features

The algorithm for determining the domain architecture of all proteins within MiST is as follows:

1. Create a FASTA file containing all the proteins for those components that have not been hammered.
2. Using the hmmpfam program, search the component FASTA file against Pfam and SMART.
3. Evaluate the significance of each domain prediction and remove insignificant hits.
4. Insert the domain predictions into the Derived data section of MiST.
5. Update the Proteins table to reflect any predicted domains and construct a character string representation of the domain architecture for each protein.

Custom Perl scripts handle the majority of these tasks by "gluing" together the various stages as well as handling all database requests using the DBI module. Due to the large number of protein sequences (roughly one million) and domain models (roughly 7500) to be searched, step three is executed on our local Linux cluster using an "embarrassingly parallel" approach. This involves each slave node searching a proportional slice of the sequence space using a local copy of the HMMER software package and the Pfam and SMART domain libraries. On moderate hardware, searching a single sequence for Pfam domains requires approximately four seconds. On a single desktop server, searching all sequences against all domain models would require more than six months of continuous computation; however, we are able to accomplish this goal in just under a month with our Linux cluster. Each domain is assessed for significance using the respective scoring

scheme from each domain database. Throughout this processing pipeline, the results are exchanged in XML before being copied into MiST. The Proteins table is then updated by assigning to each protein a non-redundant list of domains it contains and a corresponding string representation of its domain architecture. This non-redundant domain list optimizes searching for proteins with specific domains. The string architecture consists of listing each domain prediction in order of its occurrence in the protein sequence and using a tilde ("~") character for each amino acid that occurs outside a domain prediction. By virtue of the string design, complicated queries may be constructed which include constraints such as a specific sequential ordering of domains and limits regarding the length of amino acids between domains. For example, to select all proteins with an N-terminal REC domain and an unknown C-terminal domain of at least 80 amino acids one could use the following regular expression: "^~{0,70}<REC>.*~{80,}$". Signal peptides, transmembrane regions, coiled-coils, and segments of low-complexity are predicted in a similar fashion, except these predictions are done locally on the server rather than on the cluster.

Step 3: Identifying and classifying signal transduction proteins

Identifying signal transduction proteins entails querying each protein's domain architecture for the presence of conserved signaling domains. We designated 118 Pfam and SMART domains as signaling domains based on the following: a) their function, b) association with other domains, c) COGs, and d) experience working with signal transduction proteins (Table 3.1). These domains were further classified as input, output, transmitter, or receiver domains and inserted into MiST. Each protein is searched for these domains and inserted into the SignalProteins table if they contain an output,

transmitter, or receiver domain. Input domains often are found in pathways other than signal transduction (e.g. metabolic pathways), and therefore proteins identified solely from an input domain are not classified as belonging to signal transduction. During this process, the Perl script also filters out various domain combinations that indicate a role other than signal transduction. The Perl scripts that constitute this pipeline may be executed on a regular schedule such that new genomes are seamlessly and automatically integrated into the database, and signal transduction blueprints for these organisms are automatically generated.

Table 3.1 Pfam and SMART domains used for identifying signal transduction proteins.

Domain	Source	Classification	Function
BLUF	Pfam	Input	Cofactor binding
FeS	Pfam	Input	Cofactor binding
Fer4	Pfam	Input	Cofactor binding
Hemerythrin	Pfam	Input	Cofactor binding
HhH-GPD	Pfam	Input	Cofactor binding
NIR_SIR	Pfam	Input	Cofactor binding
NIR_SIR_ferr	Pfam	Input	Cofactor binding
Nitro_FeMo-Co	Pfam	Input	Cofactor binding
Phytochrome	Pfam	Input	Cofactor binding
Aminotran_1_2	Pfam	Input	Enzymatic
Arch_ATPase	Pfam	Input	Enzymatic
Citrate_synt	Pfam	Input	Enzymatic
Cyanate_lyase	Pfam	Input	Enzymatic
EPSP_synthase	Pfam	Input	Enzymatic
FmdA_AmdA	Pfam	Input	Enzymatic
GATase_2	Pfam	Input	Enzymatic
Glucokinase	Pfam	Input	Enzymatic
Glycos_trans_3N	Pfam	Input	Enzymatic
Glyoxalase	Pfam	Input	Enzymatic
HEAT_PBS	Pfam	Input	Enzymatic
HEM4	Pfam	Input	Enzymatic
NTP_transf_2	Pfam	Input	Enzymatic
NUDIX	Pfam	Input	Enzymatic
Nitroreductase	Pfam	Input	Enzymatic
PALP	Pfam	Input	Enzymatic
PTS-HPr	Pfam	Input	Enzymatic
PTS_EIIC	Pfam	Input	Enzymatic
Peptidase_M23	Pfam	Input	Enzymatic
PfkB	Pfam	Input	Enzymatic
Pribosyltran	Pfam	Input	Enzymatic
Pyr_redox	Pfam	Input	Enzymatic
Rhodanese	Pfam	Input	Enzymatic
SKI	Pfam	Input	Enzymatic
peroxidase	Pfam	Input	Enzymatic
CBS	Pfam	Input	Protein-protein interaction
HAMP	Pfam	Input	Protein-protein interaction
TPR_1	Pfam	Input	Protein-protein interaction
TPR_2	Pfam	Input	Protein-protein interaction
TPR_3	Pfam	Input	Protein-protein interaction
TPR_4	Pfam	Input	Protein-protein interaction
ACT	Pfam	Input	Small-molecule binding
Ada_Zn_binding	Pfam	Input	Small-molecule binding
AlkA_N	Pfam	Input	Small-molecule binding
AraC_binding	Pfam	Input	Small-molecule binding
Autoind_bind	Pfam	Input	Small-molecule binding
CHASE	Pfam	Input	Small-molecule binding
Cache	Pfam	Input	Small-molecule binding
Diacid_rec	Pfam	Input	Small-molecule binding
FHA	Pfam	Input	Small-molecule binding
Fe_dep_repr_C	Pfam	Input	Small-molecule binding
FeoA	Pfam	Input	Small-molecule binding
GAF	Pfam	Input	Small-molecule binding
HMA	Pfam	Input	Small-molecule binding
LysR_substrate	Pfam	Input	Small-molecule binding
PAS	Pfam	Input	Small-molecule binding
PAS	SMART	Input	Small-molecule binding
PAC	SMART	Input	Small-molecule binding
PBP	Pfam	Input	Small-molecule binding
PBP	Pfam	Input	Small-molecule binding

Table 3.1 (continued)

PBP	Pfam	Input	Small-molecule binding
SIS	Pfam	Input	Small-molecule binding
STAS	Pfam	Input	Small-molecule binding
TOBE	Pfam	Input	Small-molecule binding
TetR_C	Pfam	Input	Small-molecule binding
V4R	Pfam	Input	Small-molecule binding
cNMP_binding	Pfam	Input	Small-molecule binding
CHASE2	Pfam	Input	Unknown function
CHASE3	Pfam	Input	Unknown function
CHASE4	Pfam	Input	Unknown function
MASE1	Pfam	Input	Unknown function
MASE2	Pfam	Input	Unknown function
MHYT	Pfam	Input	Unknown function
TrkA_C	Pfam	Input	Unknown function
Arc	Pfam	Output	DNA-binding
Arg_repressor	Pfam	Output	DNA-binding
AsnC_trans_reg	Pfam	Output	DNA-binding
Crp	Pfam	Output	DNA-binding
CtsR	Pfam	Output	DNA-binding
DeoR	Pfam	Output	DNA-binding
Fe_dep_repress	Pfam	Output	DNA-binding
GerE	Pfam	Output	DNA-binding
GntR	Pfam	Output	DNA-binding
HTH_1	Pfam	Output	DNA-binding
HTH_10	Pfam	Output	DNA-binding
HTH_3	Pfam	Output	DNA-binding
HTH_5	Pfam	Output	DNA-binding
HTH_6	Pfam	Output	DNA-binding
HTH_7	Pfam	Output	DNA-binding
HTH_8	Pfam	Output	DNA-binding
HTH_AraC	Pfam	Output	DNA-binding
IclR	Pfam	Output	DNA-binding
LacI	Pfam	Output	DNA-binding
LytTR	Pfam	Output	DNA-binding
MarR	Pfam	Output	DNA-binding
MerR	Pfam	Output	DNA-binding
PadR	Pfam	Output	DNA-binding
RHH_1	Pfam	Output	DNA-binding
ROS_MUCR	Pfam	Output	DNA-binding
TetR_N	Pfam	Output	DNA-binding
Trans_reg_C	Pfam	Output	DNA-binding
EAL	Pfam	Output	Di-guanylate cyclase
GGDEF	Pfam	Output	Di-guanylate cyclase
HD	Pfam	Output	Hydrolase
Guanylate_cyc	Pfam	Output	Other
LytR_cpsA_psr	Pfam	Output	Other
Rrf2	Pfam	Output	Other
RseA_N	Pfam	Output	Other
PP2C_SIG	SMART	Output	Phosphatase
Pkinase	Pfam	Output	Protein kinase
ANTAR	Pfam	Output	RNA-binding
CsrA	Pfam	Output	RNA-binding
Response_reg	Pfam	Receiver	Response regulator
HATPase_c	Pfam	Transmitter	Histidine kinase
HATPase_c	SMART	Transmitter	Histidine kinase
HisKA	Pfam	Transmitter	Histidine kinase
HisKA	SMART	Transmitter	Histidine kinase
MCPsignal	Pfam	Transmitter	MCP
MA	SMART	Transmitter	MCP

Exploratory Knowledge Environment

In order to provide an effective means for exploring and utilizing the bioinformatics platform beyond local data mining, we implemented a web-based knowledge environment. The majority of programs comprising the bioinformatics platform lack graphical user-interfaces and are invoked from the command-line interface of a computer shell. While this approach typically provides the user with the most options and capabilities, the non user-friendly terminal often acts as a barrier to research for biologists – the majority of which are not comfortable or familiar with this mode of operation. To alleviate this problem, we built a user-friendly, web interface that enables users to perform basic bioinformatics related tasks as well as query and interact with the MiST database.

BLAST, HMMER, Consensus, and an alignment-shading program were implemented as part of the knowledge environment. To provide scalability and boost performance, BLAST and HMMER were remotely executed on the Linux cluster via Python scripts using the Twisted network library. The typical path of a BLAST or HMMER job is as follows:

1) The user inputs any sequence(s) to be searched via the website,
2) CGI scripts encode this request in XML and transmit the job request to the cluster,
3) the request is decoded and executed,
4) the program output is encoded into XML and transmitted back to the CGI script on the Apache web server, and finally,
5) the XML results are appropriately presented to the user.

A CGI wrapper program (Figures 3.3, 3.4) manages PSI-BLAST requests with a number of additional, useful features that include the following: customized XML-encoded input/output to PSI-BLAST, the ability to save and load searches at any iteration, download sequences from selected hits, view the domain architectures for any selected hits, and an improved visualization and user-interface. The front-end to HMMER enables the user to search multiple sequences against either Pfam or SMART and visualizes each protein's domain architecture. Consensus and the alignment shader are executed on the web server as CGI scripts. Screenshots of each of these programs in action are provided in Figures 3.5-3.8.

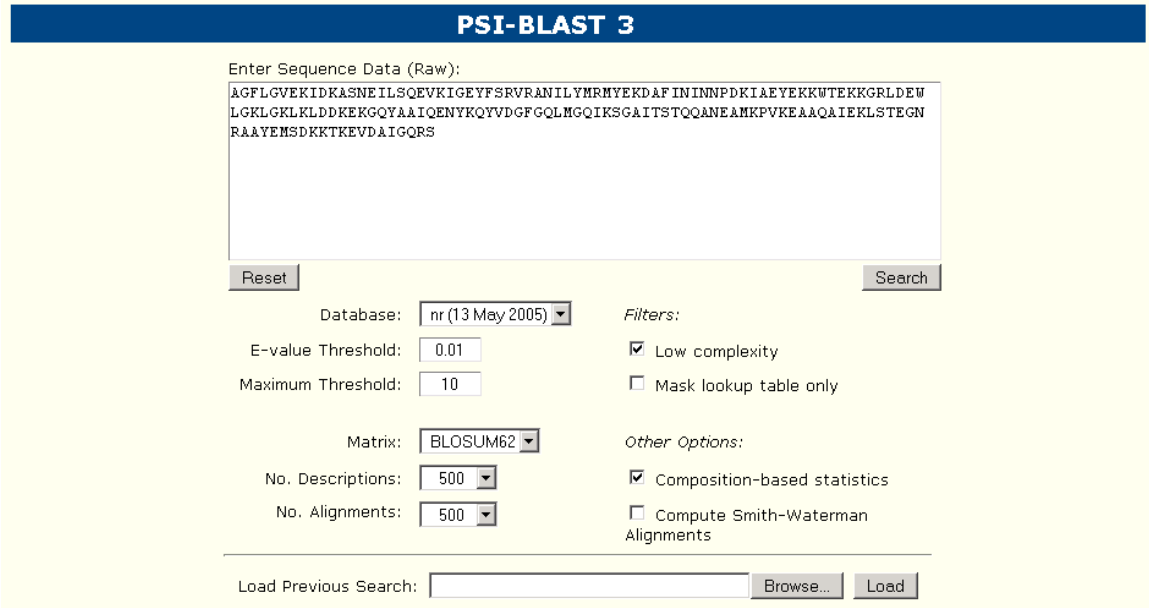


Figure 3.3 Web interface to PSI-BLAST. Input spaces are provided for the query sequence, database to search, and other various options. Users may also load previous searches using this page.

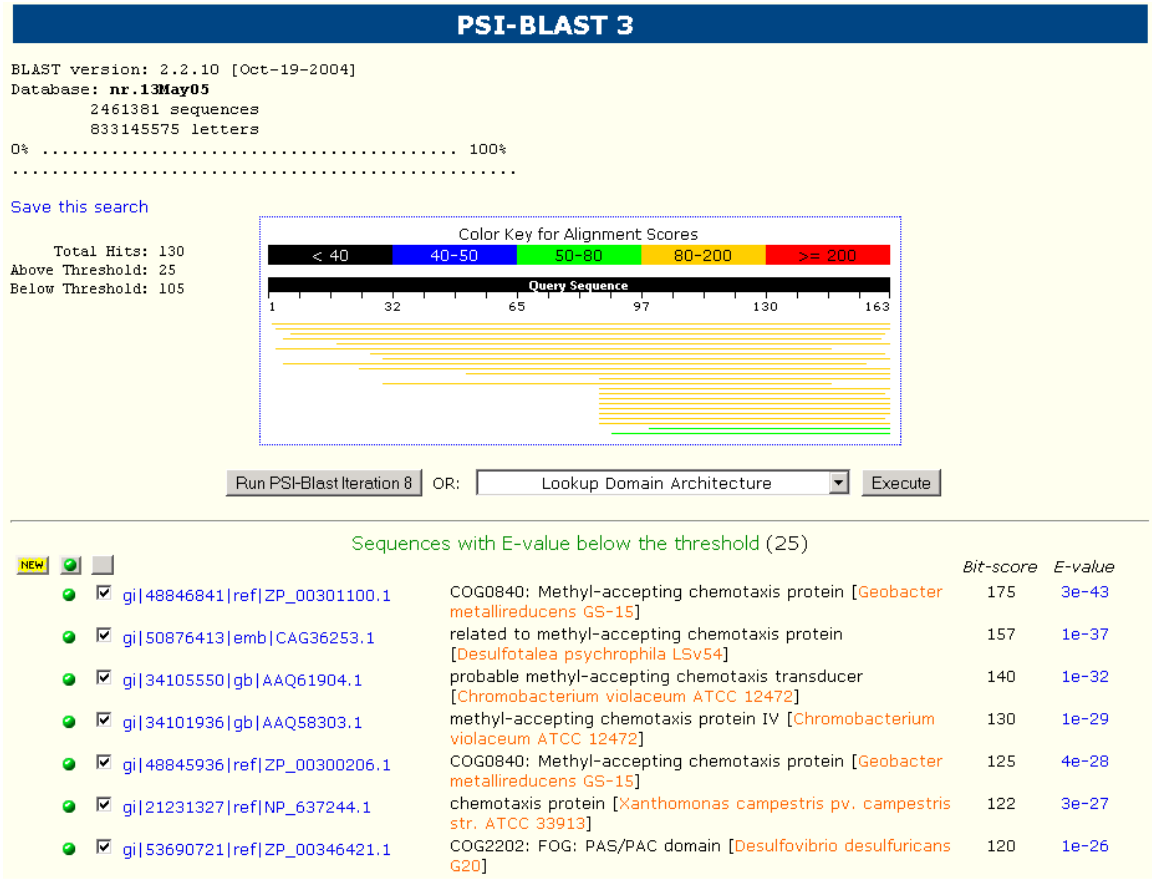


Figure 3.4 Output of results from a web-based PSI-BLAST search. Descriptive information about the search includes the BLAST version, database size, and number of hits. A graphical output displays hits to the query sequence as horizontal lines and the color of each line indicates its score. The search may be continued additional iterations or the results from this search may be used in another analysis (e.g. view the domain architectures for the selected sequences). Significant BLAST hits and information about each pairwise alignment are displayed below the graphical overview.

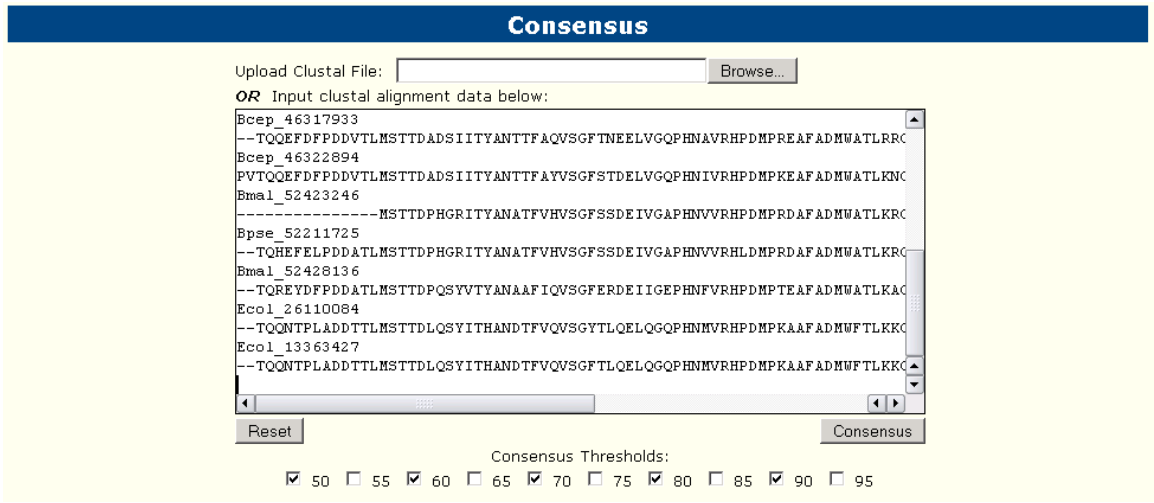


Figure 3.5 Web interface to the Consensus tool. Users upload ClustalW alignments or input the alignment directly into the provided text area. Checkboxes specify the consensus levels to produce.

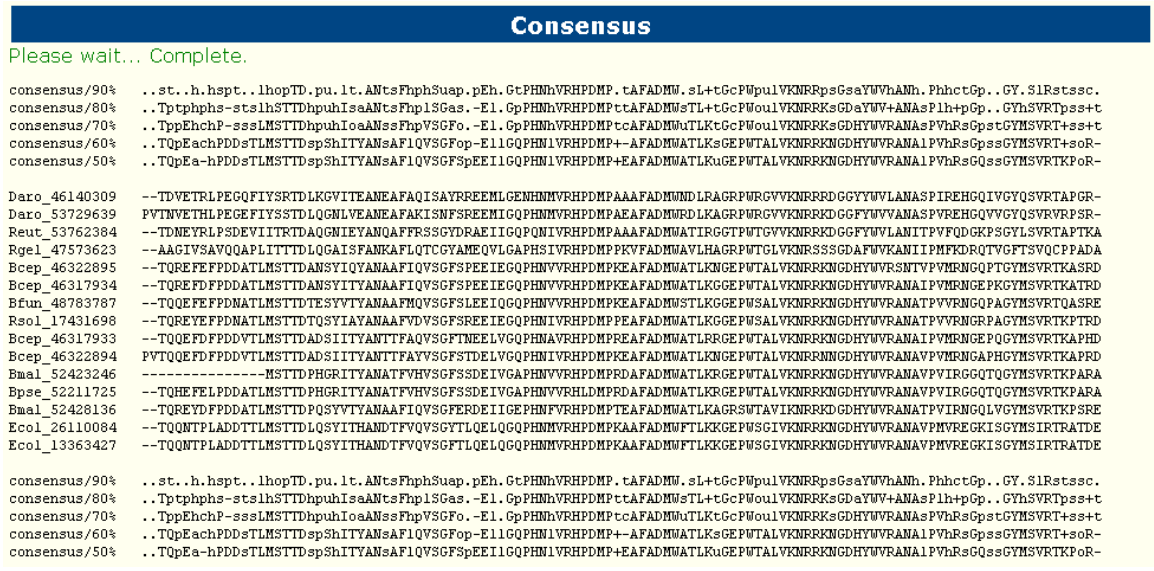


Figure 3.6 Output of results from the web-based Consensus tool. The alignment is redisplayed with the consensus above and below.

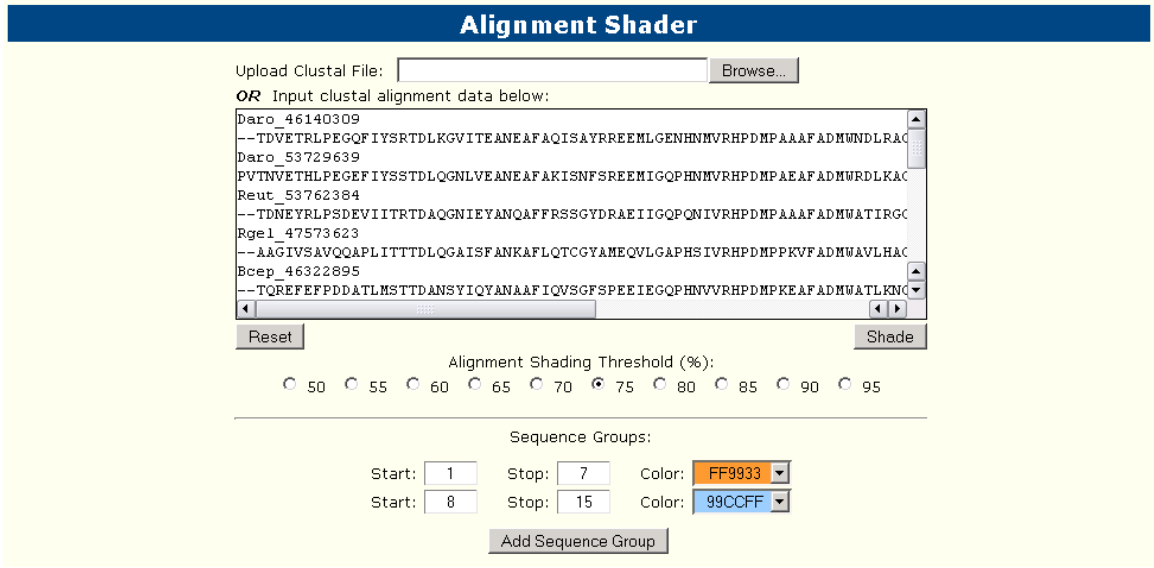


Figure 3.7 Web interface to the Alignment Shader tool. Users upload ClustalW alignments or input the alignment directly into the provided text area. The user then describes the groups of sequences within the alignment, what color to shade them, and at what similarity threshold a column must surpass for it to be shaded.

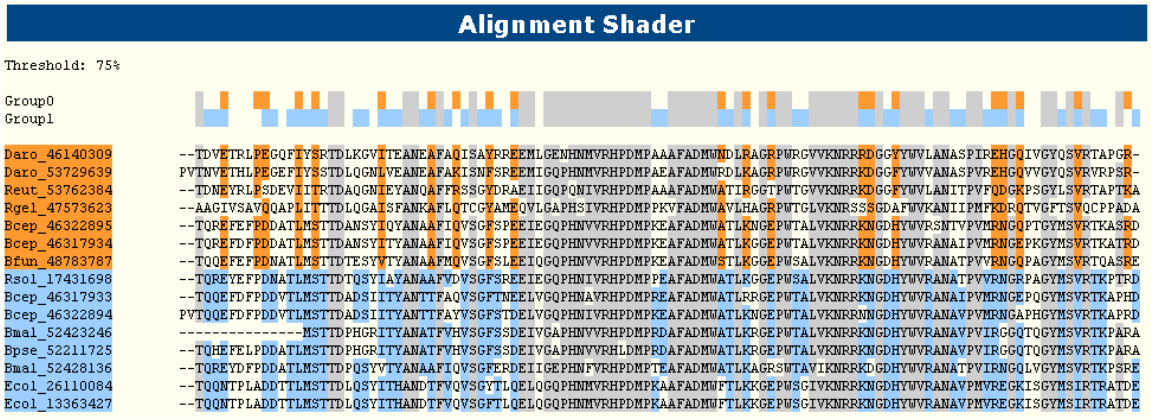


Figure 3.8 Output of results from the web-based Alignment Shader tool.

A sophisticated web interface has also been built for exploring the MiST database. Initially, the user is presented with a list of genomes in MiST organized by an adjustable level of taxonomy (Figure 3.9). Organisms of interest may be selected by clicking the checkboxes beside the organism name and then pushing the 'Select Organisms' button. This action presents the analysis selection page (Figure 3.10), which provides the ability to query the selected organisms by domain or domain architecture, description, GI, and MiST identifier. Additionally, the user may view the number of domains in each of the selected organisms by inputting a list of comma-separated domains in the 'Domain counts' field and clicking 'Count'. Clicking on an organism name displays both general and specific information about signal transduction for that organism including: descriptive data about the genome and proteome (e.g. size, chromosomes, nucleotide frequencies, etc.), a graphical representation of its signal transduction profile, querying options, and lists of the one- and two-component systems identified in this microbe (Figure 3.11). The profile presents an overall view of the number of signaling systems separated into functional categories. For example, *E. coli* contains 221 DNA-binding domains, clearly indicating a substantial level of gene regulation (Figure 3.12).

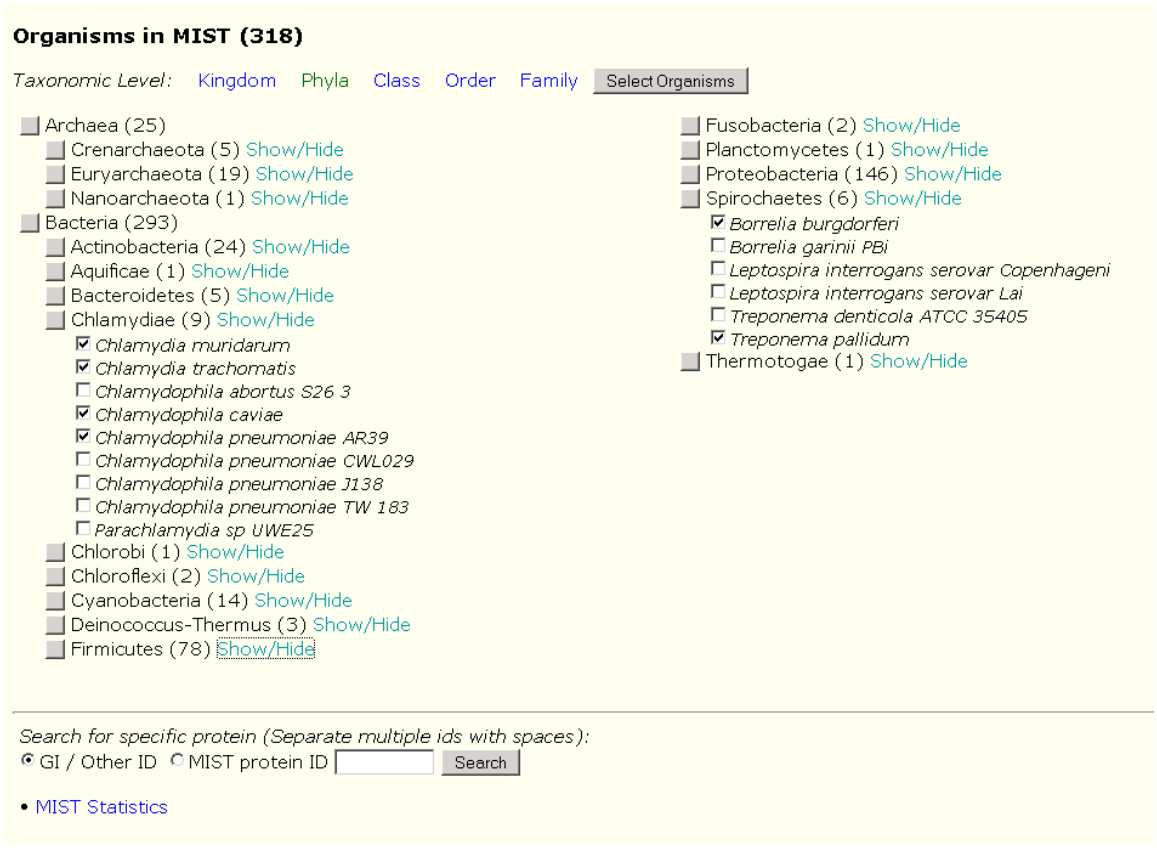


Figure 3.9 Entry web page to the MiST database. Available genomes are listed according to their taxonomy. Clicking the individual checkboxes beside each organism name selects them for further investigation. Users may also search by GI number or MiST identifiers for a specific protein.

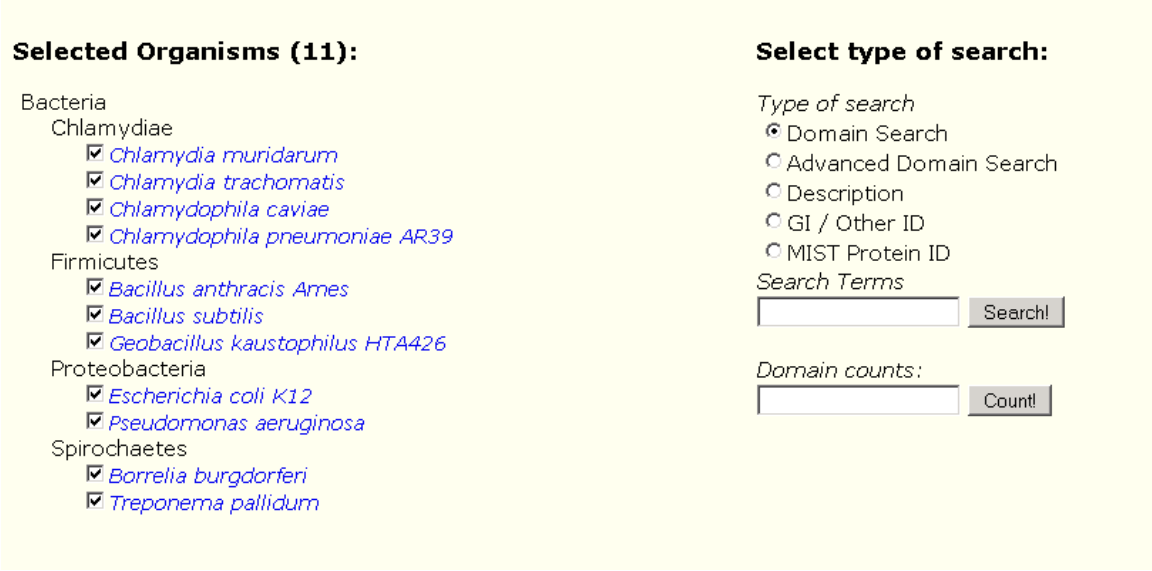


Figure 3.10 Analysis selection web page for the MiST database. Each of the previously selected organisms may be searched by domain, description, GI, or MiST identifiers. Clicking on an individual organism name displays various information for that particular organism.

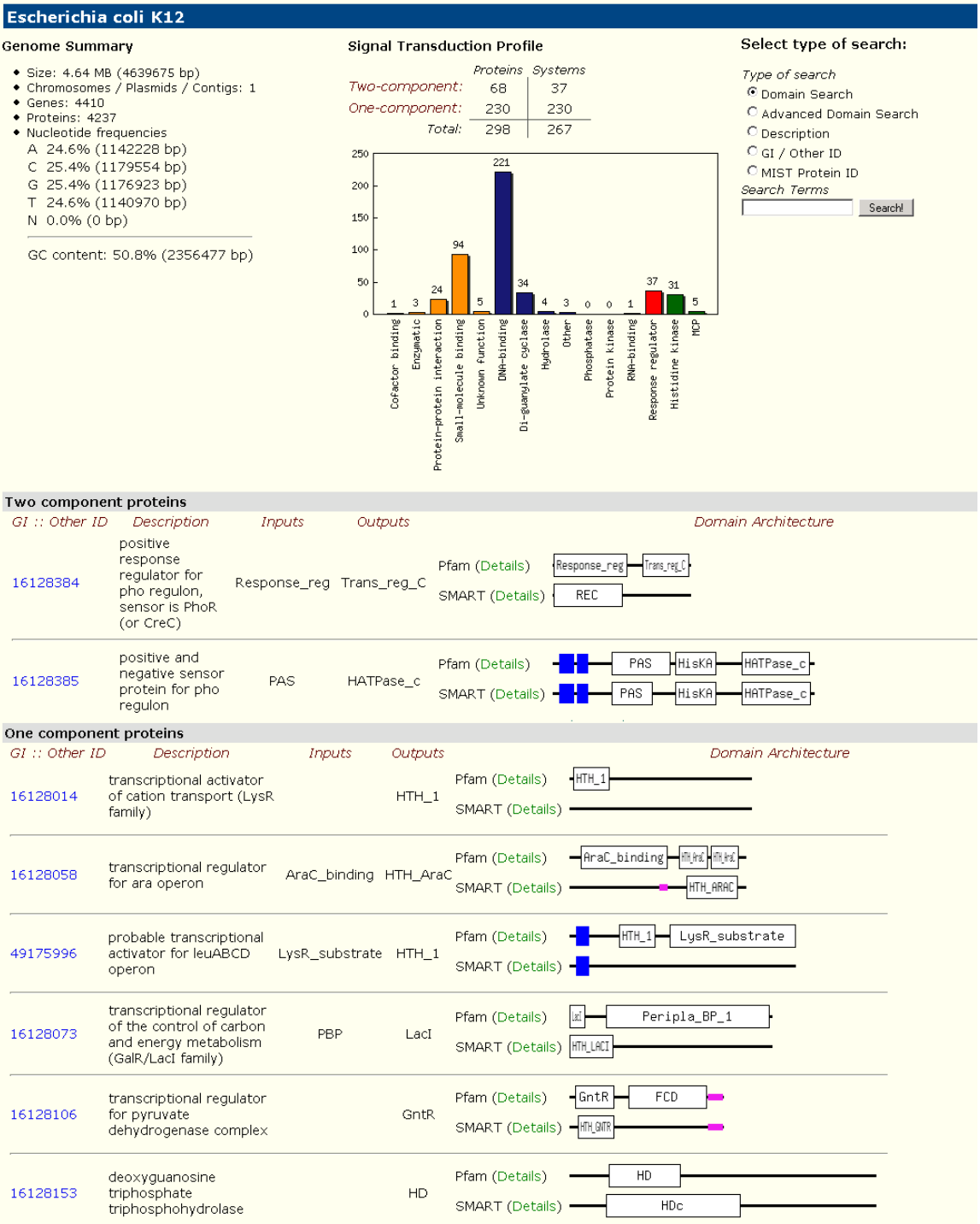


Figure 3.11 Organism specific page for the MiST database containing descriptive information about the genome, signal transduction profile, querying options, and lists of one- and two-component proteins found in this organism.

From the organism page, clicking on an individual protein hyperlink such as GI number 16128385 of the *E. coli* organism page, displays an information-rich protein/gene page (Figure 3.12). One may view basic information about this gene and protein including their RefSeq annotations, predicted Pfam and SMART domains, and the genome neighborhood. The amino acid and DNA sequences may be retrieved from the 'Sequence' hyperlinks. Additionally, one may extract a user-specified length of upstream or downstream DNA from the gene. The domain architecture section presents a linear visualization of the protein's domain architecture and other secondary features. Any predicted domains are displayed as white boxes with the domain name printed inside this box. Other secondary features are visualized by filled rectangles with the following colors: signal peptides, red; transmembrane regions, blue; coiled-coils, green; and low-complexity regions, purple. The chromosome view reveals any neighboring genes. The currently selected gene is shown in blue and other surrounding genes are colored in gray. Each gene is clickable and will display a similar page for that gene.

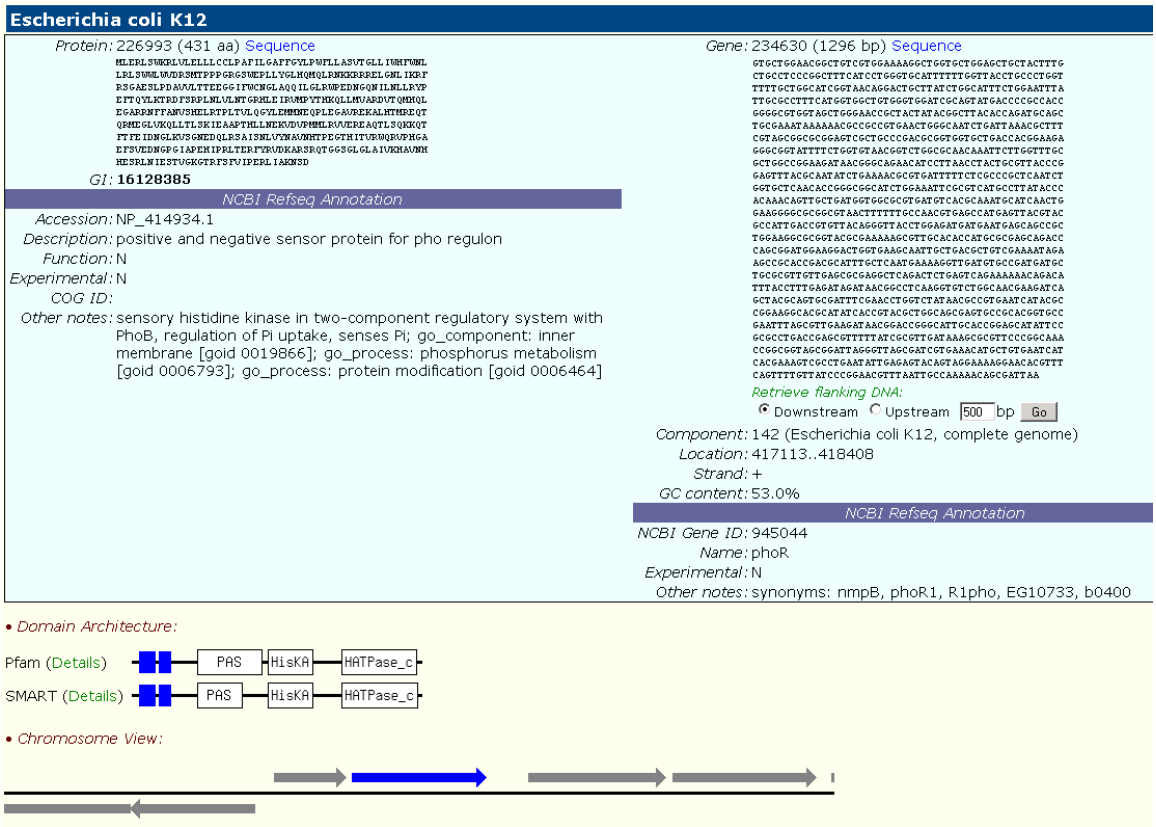


Figure 3.12 Protein and gene web page for the MiST database. Basic annotation and sequence data for a protein and its corresponding gene are displayed along with the predicted domain architecture and genome neighborhood.

CHAPTER 4

ONE-COMPONENT REGULATORS DOMINATE SIGNAL TRANSDUCTION IN PROKARYOTES

This chapter is an adapted reproduction of our recent publication in *Trends in Microbiology*, One-component Regulators Dominate Signal Transduction in Prokaryotes (2005) by Ulrich, L.E., Koonin, E.V., and Zhulin, I.B.

Introduction

Two-component systems that link environmental signals to cellular responses are viewed as the primary mode of signal transduction in prokaryotes. By analyzing information encoded in 145 prokaryotic genomes, we found that the majority of signal transduction systems consist of a single protein containing input and output domains but lacking phosphotransfer domains typical of two-component systems. One-component regulators are evolutionary older, more widely distributed among bacteria and archaea, and display a greater diversity of domains than two-component systems.

Signal transduction pathways in prokaryotes regulate cellular functions in response to environmental cues. According to the current view, prokaryotic signal transduction is comprised mostly of two-component regulatory systems which function through phosphotransfer between two key proteins, a sensor histidine kinase and a response regulator (Hoch, 2000; Hoch and Silhavy, 1995; Inouye and Dutta, 2003; Parkinson, 1993; Parkinson and Kofoid, 1992; Stock, et al., 2000). Most experimentally studied histidine kinases are membrane-bound and have an extracellular input domain,

whereas all response regulators are cytosolic. In a typical two-component system, the input domain of the sensor histidine kinase detects the environmental signal (usually a small molecule ligand) thereby activating the histidine kinase domain which autophosphorylates at a specific histidine residue (H) (Figure 4.1a). The phosphoryl group (P) is then transferred to a specific aspartate residue (D) in the receiver domain of a response regulator. Phosphorylation of the response regulator activates the output domain, which initiates the corresponding cellular response. The majority of experimentally characterized two-component systems regulate gene expression at the level of transcription via the DNA-binding helix-turn-helix (HTH) output domains of the response regulators (Hoch, 2000; Stock, et al., 2000). In addition, some response regulators contain enzymatic output domains, such as the di-guanylate cyclase (Paul, et al., 2004).

What is a One-component System?

The modular design of two-component systems is elegant, but does not seem to be the simplest possible solution to the signal transduction challenge – linking environmental stimuli to adaptive responses. A much simpler design of a signal transducer is direct fusion of an input domain to an output domain in a single protein molecule (Figure 4.1A). Indeed, it is well known that transcription of prokaryotic operons is typically regulated by single-molecule repressors and activators which, like most two-component systems, contain a ligand-binding domain and a DNA-binding HTH domain. The LacI lactose operon repressor (Lewis, et al., 1996) and the CAP activator (Kolb, et al., 1993) of *E. coli* are classic examples of such transcriptional regulators. Although these transcriptional regulators are not normally described as signal transduction systems,

we and others noticed that they contain some of the same input and output domains that are typical of two-component signal transduction systems (Anantharaman, et al., 2001; Galperin, et al., 2001; Shu, et al., 2003; Shu and Zhulin, 2002; Taylor and Zhulin, 1999; Zhulin, et al., 2003). For example, PAS (Taylor and Zhulin, 1999) and HTH are input and output domains, respectively, in the two-component system NtrB/NtrC (Weiss, et al., 2002) and in the single-molecule transcriptional regulator RocR (Calogero, et al., 1994) (Figure 4.1B). Using domain database searches, we identified many other combinations of input and output domains as direct fusions in known or predicted regulatory proteins (Figure 4.1B). Thus, one-component systems, which are defined as proteins that contain known or predicted input and output domains but lack histidine kinase and receiver domains, appear to have a repertoire of input and output domains similar to that of two-component systems and therefore might detect similar stimuli and elicit similar cellular responses. Sensory and regulatory properties of some of the one-component systems have been well documented experimentally (Shelver, et al., 1997; Spiro and Guest, 1990; Vannini, et al., 2002).

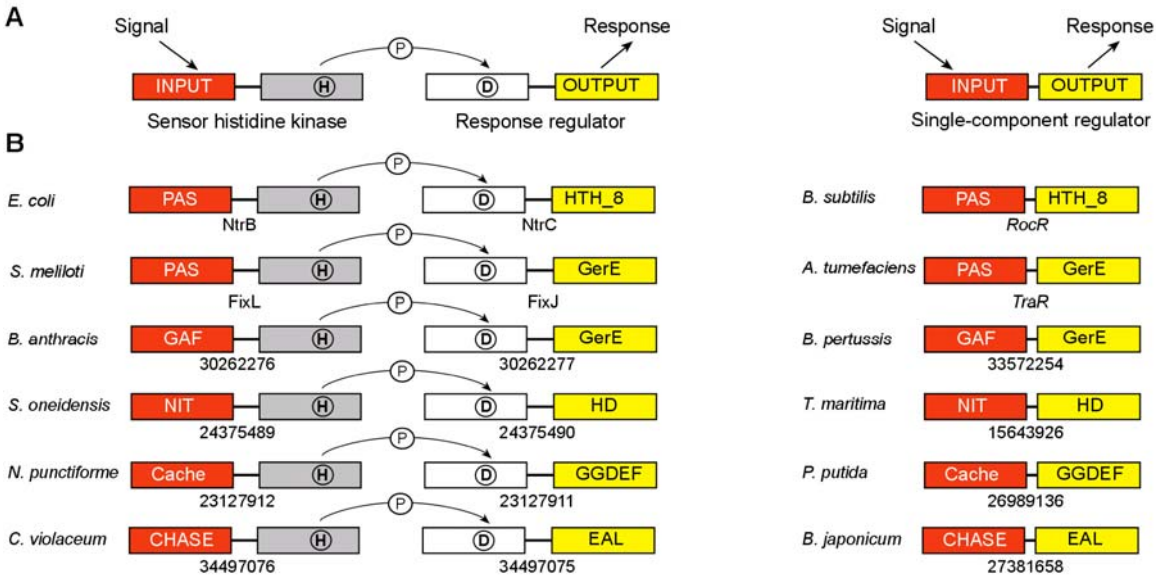


Figure 4.1 Two-component and one-component signal transduction. (**A**) A prototypical two-component signal transduction system contains input (colored red) and output (colored yellow) domains in two different proteins that communicate via a His-Asp phosphotransfer. A one-component system is a protein that contains input and output domains but lacks His-Asp phosphotransfer domains (colored gray and white). (**B**) Examples of two-component and one-component systems that utilize the same type of input and output domains. Experimentally studied proteins are identified by name, while proteins predicted from genome sequences are identified by their GenBank ID number.

Detection of Signal Transduction Proteins in Sequenced Genomes

Identification of signal transduction proteins is based on the computational domain analysis of protein sequences. We obtained 145 complete and draft prokaryotic genomes from the National Center for Biotechnology Information (a complete list of genomes is available on our website (http://genomics.biology.gatech.edu/research/TIM). Protein sequences encoded in each genome were searched against the Pfam (Bateman, et al., 2004) and SMART (Letunic, et al., 2004) domain libraries (hidden Markov models)

using the HMMER software package (http://hmmer.wustl.edu/) on a parallel Linux cluster. The resulting domain architectures were stored in a MySQL database. Custom Perl scripts were developed to query the database for domains and domain combinations using regular expressions.

Definitions of input and output domains of prokaryotic signal transduction and gene regulation categories are based on curated assignments in the Pfam-A (Bateman, et al., 2004), SMART (Letunic, et al., 2004) and Clusters of Orthologous Groups (COGs) (Tatusov, et al., 2003) resources, and recent genomic surveys (Anantharaman, et al., 2001; Galperin, et al., 2001; Zhulin, et al., 2003). A complete list of input and output domains used in this study can be found on our website (http://genomics.biology.gatech.edu/research/TIM). Two-component systems were identified by the presence of the histidine kinase (HATPase_c) and response regulator (response_reg) domains. One-component regulators were identified by the presence of one or more known output domains and the absence of histidine kinase and response regulator domains. Because many input domains involved in prokaryotic signal transduction also participate in other cellular processes (e.g., ligand-binding in transport and metabolism), they were counted only when found in a combination with the histidine kinase (two-component systems) or a known output domain (one-component regulators). In contrast, output domains typically have a single, specific regulatory function (e.g., transcriptional regulation via binding to specific promoters) and therefore all output domains were counted in the domain analysis.

One-component Versus Two-component Systems: a Survey of Bacterial and Archaeal Genomes

Exhaustive database searches and analysis (see previous section) yielded detailed information on the distribution and co-occurrence of input and output domains in 145 complete and draft prokaryotic genomes (see complete results on our website at http://genomics.biology.gatech.edu/research/TIM). Strikingly, this analysis detected many more one-component systems (~17,000) than two-component systems (~4,000). Moreover, one-component regulators show much greater diversity of the input and output domain repertoire than two-component systems (Figure 4.2). Many domains are found exclusively in one-component systems, whereas there are no unique input or output domains in two-component systems. The principal type of output activity in both classes of signal transduction systems is regulation of gene expression at the level of transcription: 87% of the known output domains in two-component systems and 84% in one-component regulators are DNA-binding HTH domains. The rest of the output domains are enzymes regulating the level of cyclic nucleotides and protein phosphorylation. The major input activity in both classes is small-molecule-binding: 96% of the known input domains in two-component systems and 93% in one-component regulators are various small-molecule-binding domains. The rest of the input domains are enzymatic and cofactor-containing (mostly redox-responsive) domains and domains involved in protein-protein interactions. Distinct input domains were detected in many but not all of the one-component regulators. However, current computational tools do not detect input domains in protein sequences of several one-component regulators that are

known to carry out specific sensory functions. For example, in the CueR Cu^{+} sensor/transcriptional activator, the HTH output domain is fused directly to a simple helix-loop-helix element (undetectable by current computational domain searches), which serves as a metal-sensing (input) domain (Changela, et al., 2003). Therefore, we hypothesize that most if not all one-component regulators detected in our genomic analysis via the identification of an output domain participate in various forms of prokaryotic signal transduction.

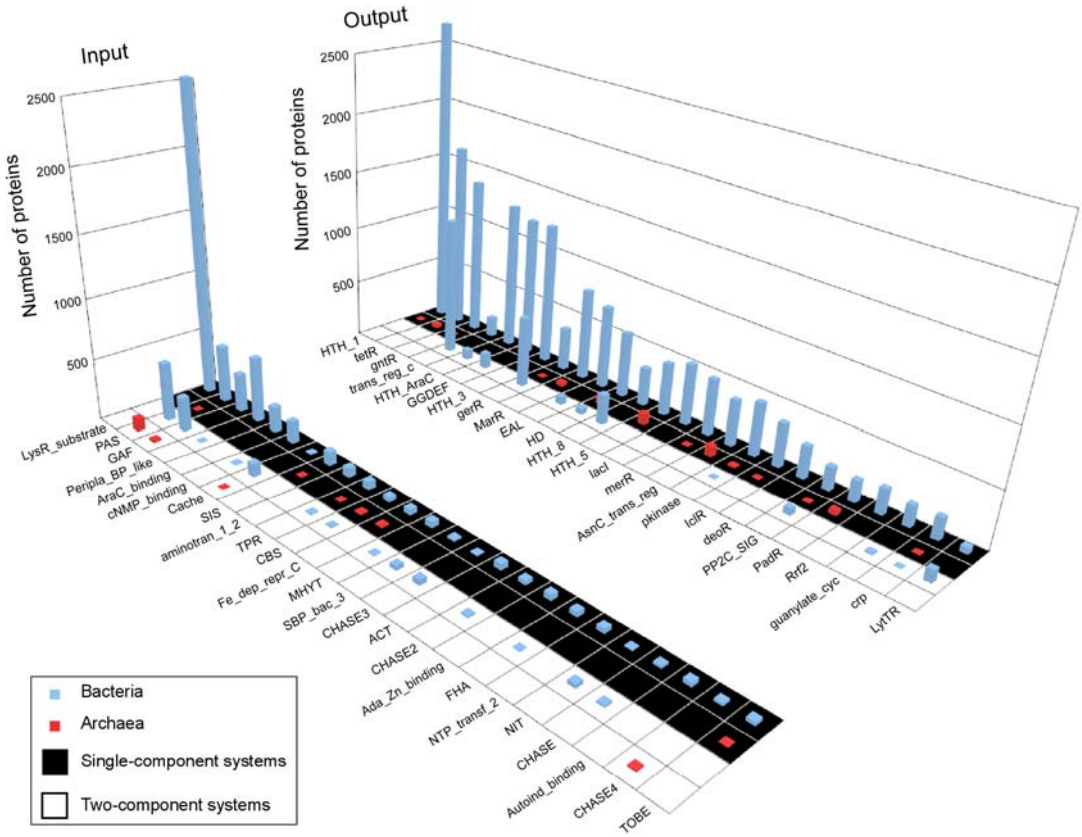


Figure 4.2 Distribution of input and output domains in bacterial and archaeal signal transduction systems. The counts of the twenty-five most abundant input and output domains in bacterial and archaeal one-component and two-component systems are shown. Domain nomenclature is from the curated Pfam-A database (see Appendix A for detailed information).

Genome Size, Lifestyle and Environment Contribute to the Complexity of Signal Transduction

The number of one-component and two-component systems per genome positively correlates with the genome size and, in both cases, is roughly proportional to the square of the total number of genes (Figure 4.3). As shown recently, signal transduction and regulation of gene expression stand out among all functional categories of proteins in showing the steepest dependence on the total number of genes (Konstantinidis and Tiedje, 2004; Van Nimwegen, 2003). This seems to reflect the disproportionate increase in the hierarchical complexity of gene regulation with the increase in genome size, which might ultimately control the maximum achievable genome size, at least in prokaryotes. The results presented here indicate that both one-component and two-component systems contribute to this increase in biological complexity, but given the similar exponents of the plots (Figure 4.3), the contribution of the more abundant one-component regulators is greater. Significant deviations from this general trend seem to reflect particular biological phenomena as well as environmental conditions in a microbial habitat. For example, the genomes of the marine cyanobacterium *Trichodesmium erythraeum* and the soil α-proteobacterium *Sinorhizobium meliloti* are comparable in size: 7.7 and 6.7 Mb, respectively. However, there are 69 one-component regulators encoded in the former (unusually few) versus 390 in the latter (unusually many). The difference in the number of two-component systems in the two genomes is much less noticeable and, in fact, not significant: 35 and 40, respectively. Both bacterial species have a versatile metabolism, which is reflected by

their large genome size. However, *S. meliloti* has a complex developmental program (Galibert, et al., 2001) and experiences significant fluctuation of various physico-chemical parameters in its microenvironments (soil, rhizosphere, plant root interior). In contrast, *T. erythraeum* does not undergo developmental changes typical of other nitrogen-fixing cyanobacteria (heterocyst formation) and lives under more or less constant environmental conditions (upper levels of tropical oceans) (Staal, et al., 2003).

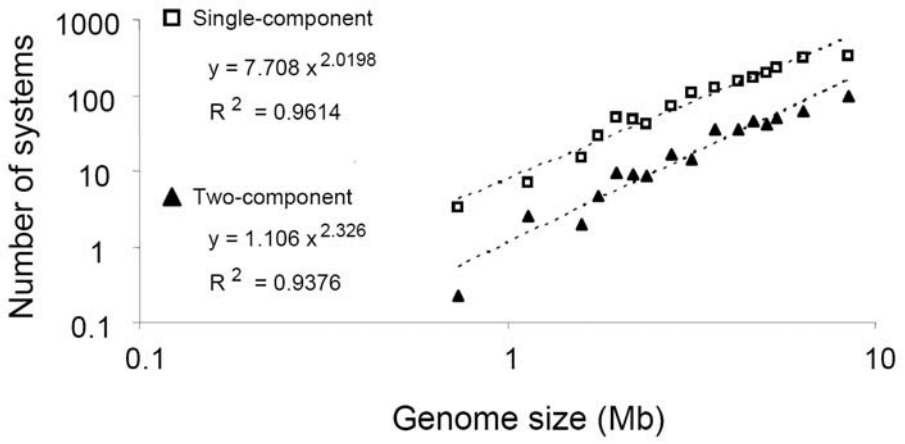


Figure 4.3 Dependence of the number of one-component and two-component signal transduction systems on the genome size. The plot is in a double logarithmic scale. One hundred forty-five genomes were ranked by size and split into 16 size classes. Each point indicates the average number of genes for one-component or two-component signal-transduction systems in the respective class.

One-component Systems as the Primordial Form of Prokaryotic Signal Transduction

Three lines of evidence suggest that one-component regulators are evolutionary precursors of the two-component systems. Firstly, the modular design of one-component systems is obviously simpler than that of two-component systems. Secondly, as shown above, the domain repertoire of one-component regulators is considerably more diverse than that of two-component systems. Finally, one-component regulators are more widely distributed among prokaryotes than two-component systems: with the exception of some parasites with highly degraded genomes, such as mycoplasmas, all prokaryotes encode a substantial diversity of one-component regulators. By contrast, two-component systems are (nearly) missing in many species, particularly, among archaea (Figure 4.2). Furthermore, archaeal two-component systems do not seem to form a coherent lineage (data not shown) and probably have been acquired from bacteria via horizontal gene transfer as previously suggested (Koretke, et al., 2000). Therefore it appears most likely that the last common ancestor of archaea and bacteria (i.e., the last common ancestor of all modern life forms) did not have two-component systems, but probably encoded several one-component regulators. Two-component systems appear to be a subsequent bacterial innovation, which emerged through insertion of histidine kinase domains and receiver domains into one-component regulators. If one-component regulatory systems comprise such a straightforward solution to the requirements of prokaryotes for signal transduction, then, what is the advantage of two-component systems? We believe that this has to do with the fact that one-component regulators detect stimuli (including environmental cues, such as gases, light, and various small molecules) almost exclusively

in the cytosol. We scanned all 25,303 protein sequences identified as components of signal transduction systems in 145 prokaryotic genomes for the presence of transmembrane regions using the DAS method (Cserzo, et al., 2004) and found that 97% of one-component regulators that contain an HTH domain do not have transmembrane regions and therefore are predicted to be cytosolic proteins. In contrast, more than 73% of the sensor histidine kinases were predicted to be membrane-associated based on the presence of one or more transmembrane regions. Thus, the fundamental difference in the sensing mode between the one-component and two-component systems is intracellular versus extracellular detection of stimuli, respectively. Extracellular sensing provides a microbe with an obvious advantage compared to exclusive intracellular sensing. However, because more than 80% of signal transduction pathways involve DNA-binding, arrangement of single-molecule regulators in the membrane would place major constraints on their ability to interact with their targets in genomic DNA. A straightforward and efficient solution to this problem is dividing the signal transduction system into two proteins, a membrane-bound sensor and a soluble, cytosolic DNA-binding regulator, which are linked via a phosphotransfer relay. Hence the emergence of the two-component signals transduction systems.

Conclusions

The availability of a large number of sequenced prokaryotic genomes allowed us to reveal the dominance of one-component signal-transduction systems and the apparent ancestor-descendant relationship between them and the two-component systems in prokaryotes. It has been noticed previously that some of the transcriptional regulators in prokaryotes possess sensory properties. However, to our knowledge, the fact that two-

component signal transduction systems utilize a subset of the input and output domains that are present in one-component regulators so far has not been recognized, and neither was the extraordinary combinatorial diversity of one-component regulators.

Acknowledgements

The literature on signal transduction in prokaryotes is vast. We extend our apologies and appreciation to all colleagues whose work is not cited here solely due to space limitations. We thank Eugene Koonin for his participation in co-authoring this manuscript. We thank Yuri Wolf for assistance with the genome size dependence analysis and Michael Galperin, Susan Golden and Sydney Kustu for helpful discussions. This work was supported by research grants GM72285 from National Institutes of Health and EIA-0219079 from National Science Foundation (to I.B.Z.). L.E.U. was supported by IGERT-0221600 grant from National Science Foundation.

CHAPTER 5

FOUR-HELIX BUNDLE: A UBIQUITOUS SENSORY MODULE IN PROKARYOTIC SIGNAL TRANSDUCTION

This chapter is an adapted reproduction of our recent publication in *Bioinformatics*, Four-Helix Bundle: a Ubiquitous Sensory module in Prokaryotic Signal Transduction (2005) by Ulrich, L.E. and Zhulin, I.B.

Abstract

Transmembrane chemoreceptors in *Escherichia coli* utilize ligand-binding domains for detecting various external signals. The structure of this domain in the *E. coli* aspartate receptor, Tar, is known and its signal transduction mechanism is under investigation. Current domain models for this important sensory module are inaccurate and therefore cannot reveal the distribution of this domain within the current genomic landscape. We carried out sensitive and exhaustive PSI-BLAST searches initiated with the sequence corresponding to a known structure of the four-helix, ligand-binding domain of the aspartate chemoreceptor. From the resulting sequences, we built a multiple sequence alignment for this domain family, which confirmed that the current TarH model is erroneous and fails to detect most of the domain homologs. In the process, we developed a technique that visualizes the secondary structure prediction of each protein sequence in order to improve the multiple sequence alignment. We found that the four-helix up-and-down bundle represents a large domain family and includes representatives

of all major classes of prokaryotic signal transduction, namely histidine kinases, di-guanylate cyclases, and chemotaxis receptors.

Introduction

E. coli detects several attractant and repellent compounds using transmembrane chemoreceptors (also known as methyl-accepting chemotaxis proteins or MCPs) that transmit information to flagellar motors via a signal transduction pathway (Falke and Hazelbauer, 2001; Sourjik, 2004). Many other motile bacteria and archaea have homologous pathways for regulating motility, and MCPs comprise a large protein superfamily (Zhulin, 2001). All MCPs have a conserved signaling domain and variable sensory domains. The structure of the sensory (ligand-binding) domain of the *E. coli* chemoreceptor, Tar, has been solved (Bowie, et al., 1995; Milburn, et al., 1991) and revealed an antiparallel four-helix bundle. Three other transmembrane chemoreceptors of *E. coli* – Tsr, Trg, and Tap – were proposed to have homologous ligand-binding domains based on limited sequence similarity (Zhulin, 2001). More recently, a good quality 3D model of the Trg ligand-binding domain has been built using the Tar crystal structure (Peach, et al., 2002). The model was consistent with previously obtained mutational data and is strongly supported by selective formation of disulfides between introduced cysteines (W. Lai and G.L. Hazelbauer, personal communication).

Both leading domain databases, Pfam (Bateman, et al., 2004) and SMART (Letunic, et al., 2004), contain a domain model called TarH (accession number SM00319), which has been built from the multiple alignment of protein sequences corresponding to the ligand-binding domains of Tar, Tsr, Trg, and Tap from *E. coli* and closely related enteric bacteria, and three sequences of predicted ligand-binding domains

from the *Bacillus subtilis* chemoreceptors McpA, McpB, and TlpB. The Pfam and SMART TarH models recognize 79 and 42 sequences in the non-redundant database, respectively. Because McpA, McpB, and TlpB contain a Cache domain (Anantharaman and Aravind, 2000) as their core ligand-binding element (which is not present in *E. coli* chemoreceptors), we questioned the appropriateness of their alignment to the *E. coli* ligand-binding domains in the TarH model. Furthermore, the TarH alignment contains several large gaps introduced by including the *B. subtilis* sequences and is of an overall poor quality.

Methods

We began the domain analysis by initiating an exhaustive and sensitive PSI-BLAST search (Altschul, et al., 1997) against the non-redundant database (May 13, 2005) with residues 34-190 of the ligand-binding domain from the *E. coli* Tar chemoreceptor. This region is flanked by two transmembrane helices and encompasses the entire solved structure (residues 39-178) of this sensory domain. PSI-BLAST searches (E-value cutoff 0.01) were continued until no new homologs were identified. Because this domain is highly alpha-helical, a search was stopped upon hits to unrelated random coils such as those found in myosin- or laminin-like proteins. Partial and duplicate sequences were excluded from this analysis.

Multiple sequence alignments were constructed using the MUSCLE (Edgar, 2004) and ClustalW (Thompson, et al., 1994) programs and manually adjusted. While editing this alignment we developed VISSA (**V**isualization of **S**econdary **S**tructure elements for **I**mproving multiple **A**lignments), a technique that involves visualization of predicted secondary structure elements (using color) on each sequence within an MSA.

This visualization allows a rapid check for consistency between the MSA features (e.g. gaps) and the predicted secondary structure. The VISSA technique consists of two steps: *data processing* and *visualization* (Figure 5.1):

1) *Data processing*: A ClustalW-formatted MSA is loaded into memory and the secondary structure of each sequence is predicted using PSIPRED (Jones, 1999). The alignment, prediction results, and various metadata are stored in an XML file. In addition to the alignment, this XML file contains the predicted secondary structure and confidence values for each sequence in the MSA.
2) *Visualization*: The results within the XML data file are then visualized by creating an HTML document, which has the background of alpha helices and beta strands shaded red and blue, respectively, in proportion to the confidence of each prediction – darker shades representing a higher confidence value. The text color of loop regions is shaded in a similar fashion.

Custom Perl scripts handle the data processing and visualization steps. The technical implementation was meant to be as abstract as possible. An XML document specification is provided which describes how to organize the data for a multiple alignment in conjunction with its secondary structure. Given this abstraction, the individual may then predict the secondary structure using any particular tool and visualize this data with VISSA provided that the data conforms to this XML specification. Similarly, others may wish to visualize this information differently or process this data with another program. In such cases, the intermediate XML document is a structured, machine and human readable data file enabling the user to easily process the data. Sequence conservation

information and the visualized secondary structure of the four-helix bundle domains enabled us to produce a much better quality alignment.

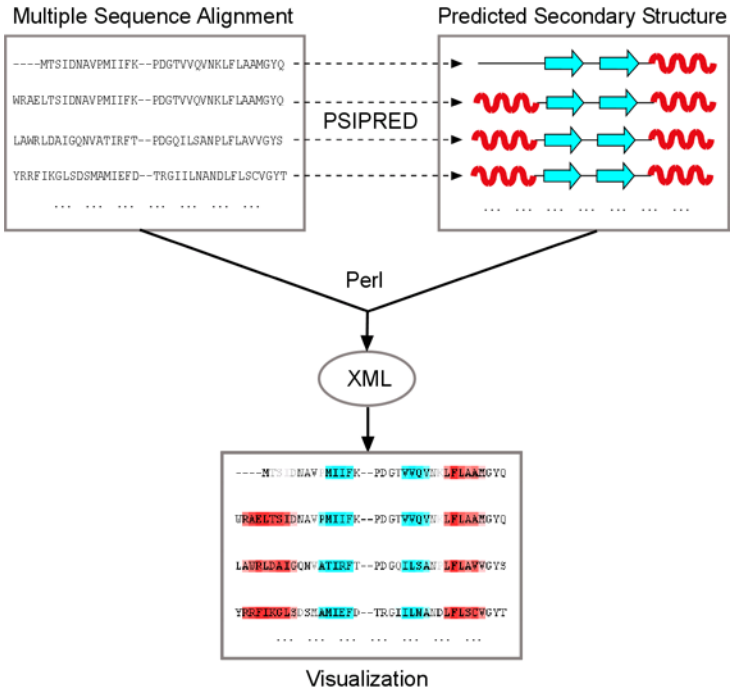


Figure 5.1 Overview of the VISSA process. For each sequence in a multiple alignment, its secondary structure is predicted and both the alignment and structural information stored in an XML document. This XML data is then visualized by coloring/shading the amino acids that correspond to the predicted secondary structure elements.

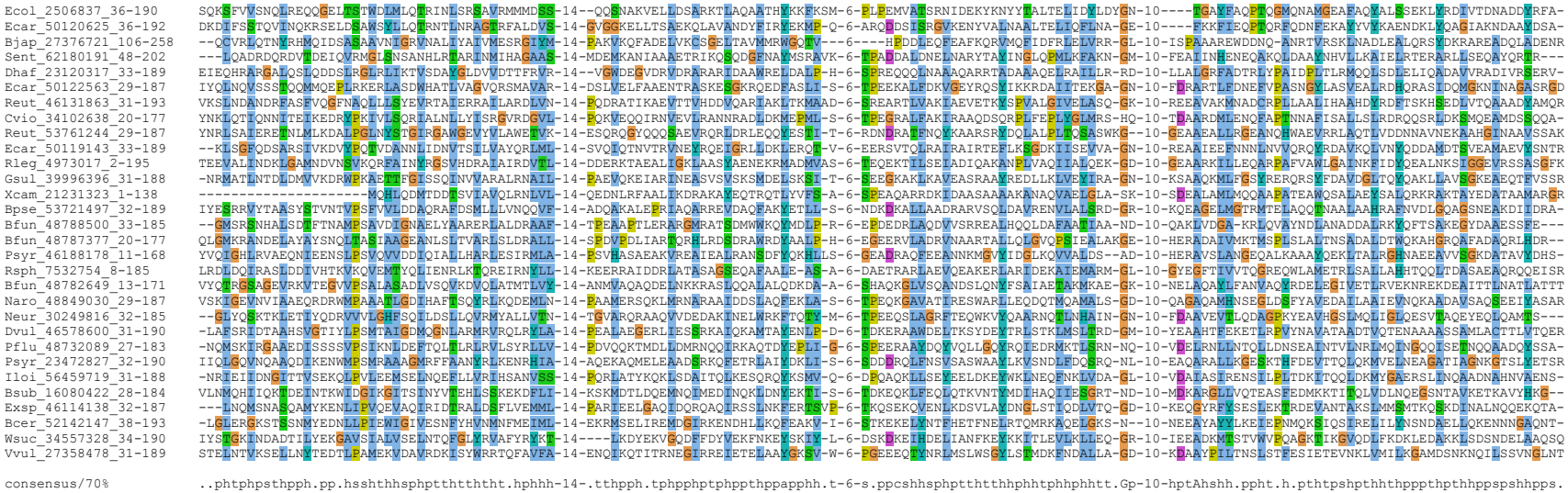
Domain architectures were predicted using the HMMER software package (Eddy, 1998) and the Pfam (version 17.0) and SMART (version 4.0) domain libraries. Signal peptides and transmembrane regions and their topology were assigned using Phobius

(Kall, et al., 2004). The secondary structure of each homolog was predicted using PSIPRED (Jones, 1999). Sequence logos based on the information content were created with the Berkeley weblogo server (Crooks, et al., 2004). The three-dimensional structure of the Tar ligand-binding domain (PDB accession 2ASR) was visualized using the PyMOL Molecular Graphic System (www.pymol.org).

Results and Discussion

Exhaustive PSI-BLAST searches resulted in the detection of more than 700 copies of the domain, which we termed 4HB_MCP for a four-helix bundle found predominantly in the N-terminal region of MCPs. These results are in striking contrast with the performance of the current TarH model in domain databases, which recognizes relatively few sequences. None of the searches detected the 4HB_MCP domain in the McpA, McpB, and TlpB proteins that were used in the seed alignment for the TarH model, thus confirming its deficiency.

A



B

Figure 5.2 Alignment of representative members of the 4HB_MCP domain. (**A**) An alignment of thirty members of the 4HB_MCP domain from the seed alignment. Conserved residues and their positions are colored with the ClustalX scheme using Jalview (Clamp, et al., 2004): orange – glycine (G), yellow – proline (P), blue – small and hydrophobic amino acids (A, V, L, I, M, F, W), green – hydroxyl and amine amino acids (S, T, N, Q), red – charged amino acids (D, E, R, K), cyan – histidine (H) and tyrosine (Y). (**B**) An alignment of the same thirty members illustrating the VISSA visualization. Regions containing predicted alpha helices and beta sheets have the background shaded red and blue, respectively. The intensity of each shade is directly proportional to the confidence of a given prediction – a darker intensity representing a higher confidence. The identifier for each sequence consists of a species abbreviation, the GenBank identifier, and the coordinates of the sequence. Species abbreviations are as follows: Bcer, *Bacillus cereus*; Bfun, *Burkholderia fungorum*; Bjap, *Bradyrhizobium japonicum*; Bpse, *Burkholderia pseudomallei*; Bsub, *Bacillus subtilis*; Cvio, *Chromobacterium violaceum*; Dhaf, *Desulfitobacterium hafniense*; Dvul, *Desulfovibrio vulgaris*; Ecar, *Erwinia carotovora*; Ecol, *Escherichia coli*; Exsp, *Exiguobacterium sp.*; Gsul, *Geobacter sulfurreducens*; Iloi, *Idiomarina loihiensis*; Naro, *Novosphingobium aromaticivorans*; Neur, *Nitrosomonas europaea*; Pflu, *Pseudomonas fluorescens*; Psyr, *Pseudomonas syringae*; Reut, *Ralstonia eutropha*; Rleg, *Rhizobium leguminosarum*; Rsph, *Rhodobacter sphaeroides*; Sent, *Salmonella enterica*; Vvul, *Vibrio vulnificus*; Xcam, *Xanthomonas campestris*; Wsuc, *Wolinella succinogenes*.

Representative and complete multiple sequence alignments of the 4HB_MCP domain were constructed and edited using the VISSA technique, which revealed several inconsistent features in the alignment (Appendix B, Figure B.1). For example, gaps have been placed inside predicted alpha helices and structural elements have been misaligned in several sequences. Efficient visualization of anomalies allowed us to quickly edit the alignment for consistency in both sequence and secondary structure conservation. VISSA readily shows the four alpha helices of this domain and the alignment is consistent with each protein's predicted secondary structure. The revised seed and complete alignments are shown in Figures B.2 and B.3 of Appendix B, respectively. The representative alignment (Figure 5.2) and its information content (Figure 5.3A) revealed several conserved residues. The most conspicuous ones are two tyrosine residues in the second and third alpha helices that appear to interact in maintaining the bundle as revealed by visualizing the corresponding positions in the Tar structure (Figure 5.3B). The overall conservation pattern of hydrophobic and polar residues reflects the coiled coil nature of this domain. No obvious conserved binding sites can be seen within the alignment, which is expected as known homologs bind very different ligands: aspartate and maltose-binding protein by Tar, serine by Tsr, ribose- and galactose-binding proteins by Trg, and dipeptides by Tap (Falke and Hazelbauer, 2001). Interestingly, the recently described CHASE3 domain family (Zhulin, et al., 2003) was found to comprise a relatively small (approximately 12%) subfamily within 4HB_MCP. The 4HB_MCP containing sensory proteins were found in many major bacterial lineages; however, it is completely missing from eukaryotes and there are only few instances of their presence in Archaea, which are

likely the result of horizontal gene transfer. This fact strongly suggests a bacterial origin for the 4HB_MCP domain.

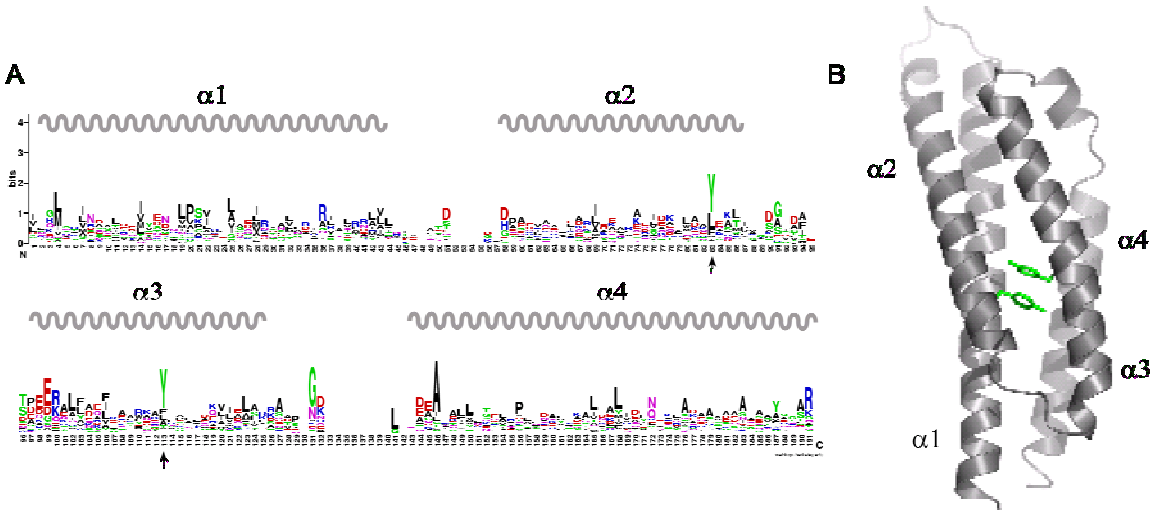


Figure 5.3 Visualization of conserved residues in the 4HB_MCP domain. (**A**) A sequence logo generated for the multiple alignment of 282 domain sequences obtained in the PSI-BLAST search initiated with the ligand-binding domain of the *E. coli* Tar chemoreceptor. The secondary structure of Tar is shown above the logo. (**B**) Conserved tyrosine residues in helices 2 and 3 interact to maintain packing of the four-helix bundle.

Domain architectures of all 4HB_MCP-containing proteins were determined using the Pfam and SMART domain libraries (Figure 5.4). The 4HB_MCP domain was found in four major classes of prokaryotic signal transduction: MCPs, sensor histidine kinases, di-guanylate cyclases and di-esterases, and guanylate/adenylate cyclases. Furthermore, this domain is found as an orphan sensory domain and in combination with

other sensory domains as an independent signal transduction module. The 4HB_MCP domain is always found between two predicted transmembrane helices, indicating that it solely detects extracellular signals. In most cases, this domain is associated with a cytoplasmic HAMP domain (Aravind and Ponting, 1999) suggesting that most 4HB_MCP proteins might share the mechanism of transmembrane signaling, which has been extensively studied in *E. coli* chemoreceptors (Ottemann, et al., 1999).

In conclusion, we have identified a large domain superfamily that plays an important role in detecting extracellular signals by bacterial transmembrane receptors involved in various signal transduction pathways. Furthermore, we have developed a visualization technique that uses secondary structure predictions to facilitate producing an accurate alignment of related protein sequences.

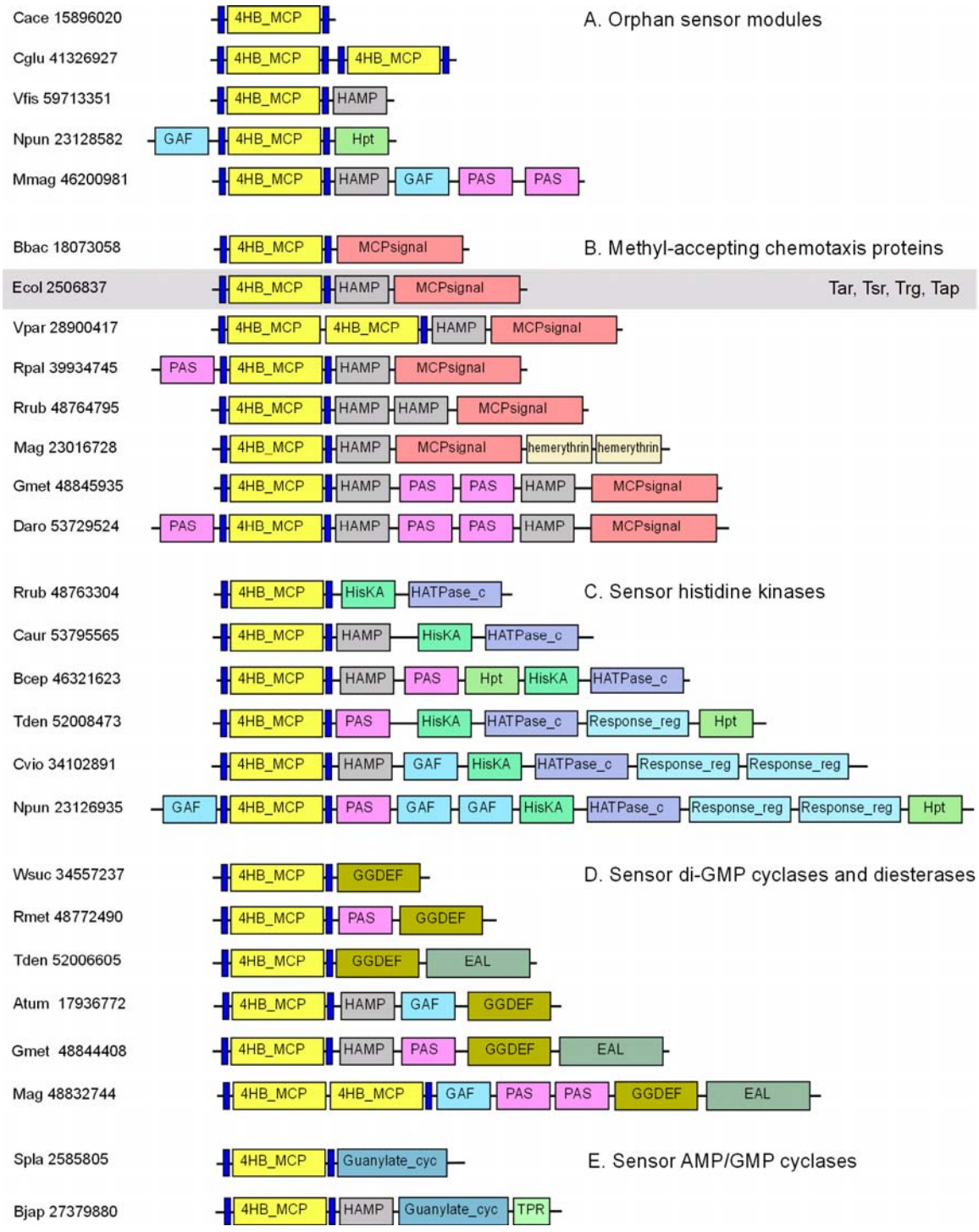


Figure 5.4 A schematic view of the domain architecture of 4HB_MCP containing proteins. Domain nomenclature is according to Pfam (Bateman, et al., 2004). Protein GenBank identifiers and the species abbreviations are shown. Species abbreviations: Atum, *Agrobacterium tumefaciens*; Bbac, *Bdellovibrio bacteriovorus*; Bcep, *Burkholderia cepacia*; Bjap, *Bradyrhizobium japonicum*; Cace, *Clostridium acetobutylicum*; Caur,

Chloroflexus aurantiacus; Cglu, *Corynebacterium glutamicum*; Cvio, *Chromobacterium violaceum*; Daro, *Dechloromonas aromatica*; Ecol, *Escherichia coli*; Gmet, *Geobacter metallireducens*; Mag, *Magnetococcus* sp.; Mmag, *Magnetospirillum magnetotacticum*; Npun, *Nostoc punctiforme*; Rmet, *Ralstonia metallidurans*; Rrub, *Rhodospirillum rubrum*; Spla, *Spiruina platensis*; Tden, *Thiobacillus denitrificans*; Vfis, *Vibrio fischeri*; Vpar, *Vibrio parahaemolyticus*; Wsuc, *Wolinella succinogenes*.

CHAPTER 6

RESOLVING THE FUNCTION OF CHEMOTAXIS PAS DOMAINS THROUGH PROTEIN SEQUENCE ANALYSIS

This chapter is ready for publication. It was recently submitted to the *Bioinformatics* journal and rejected after peer-review; however, the editor recommended addressing the reviewer's concerns and resubmitting the revised manuscript.

Abstract

PAS domains constitute a widespread superfamily of sensory modules. Current computational techniques identify PAS domains in thousands of protein sequences; however they fail to predict the specific function of a given PAS domain. We sought to design a sequence analysis-based approach that could resolve the function of chemotaxis PAS domains.

We identified 274 PAS domains found in association with chemotaxis receptors and classified them into distinct subfamilies. The PAS_Aer subfamily contains the experimentally characterized, FAD-containing redox sensor, Aer, and several critical residues for its function as a redox sensor are solely conserved in this subfamily. We predict that all PAS_Aer domains are redox sensors, whereas members of another subfamily, PAS_Che are predicted to bind a specific, but unknown ligand and represent a good target for experimental studies. Here we demonstrate that standard bioinformatics approaches may be applied to multifunctional domains in order to further resolve their function.

Introduction

Protein domains are structurally compact, independently folding units of a protein. Such structural units have identifiable patterns of amino acid conservation that can be recognized and modeled using profile hidden Markov models (HMMs) (Eddy, 1998). Built from a multiple sequence alignment of related sequences, an HMM putatively represents the structure and function of a protein domain. Many protein domains represented as HMMs are stored in the primary domain databases, Pfam (Bateman, et al., 2004) and SMART (Letunic, et al., 2004). Although HMMs capture the structural properties of a domain, it is often difficult to derive an exact biological function from general profiles that characterize large domain superfamilies. This is especially true for domains involved in signal transduction, as many of them are highly variable in sequence content and length and are associated with a broad spectrum of other domains.

One of the most diverse and functionally important domains employed in signal transduction is the PAS superfamily (Ponting and Aravind, 1997; Zhulin and Taylor, 1997). PAS (Per-ARNT-Sim) domains comprise a widespread superfamily of sensory input modules that sense light, oxygen, redox potential, small ligands, and participate in protein-protein interactions (Zhulin and Taylor, 1997). Structurally, the PAS fold (approximately 100 amino acids in length) consists of four major segments: a loosely-conserved N-terminal cap, the PAS core, a helical linker, and the beta scaffold (Pellequer, et al., 1998). The helical linker connects the PAS core and beta scaffold such that the beta strands form a central sheet surrounded by the remaining helices and loops. Altogether, these structural segments resemble a left-handed glove, which often binds a specific cofactor. In several experimental studies, the specific function of a PAS domain has been

shown to depend upon the type of a bound cofactor. For example, photo-active yellow protein (PYP) contains a 4-hydroxy-cinnamyl chromophore which is responsive to blue light (Genick, et al., 1998), FixL has a heme group coordinated to a histidine residue that is able to directly bind oxygen (Gong, et al., 1998), Aer and NifL proteins are able to detect changes in redox potential via their FAD-containing PAS domains (Macheroux, et al., 1998; Repik, et al., 2000), and plant phototropin is a light sensor utilizing an FMN-containing PAS domain (Harper, et al., 2003).

Until recently, both Pfam and SMART represented the PAS domain as two separate motifs, PAS and PAC, because of significant sequence variation in the helical linker region that was difficult to model; however, these two motifs form a single structural and functional unit and should not be considered as separate domains (Hefti, et al., 2004; Taylor and Zhulin, 1999; Taylor and Zhulin, 1999). The NR database contains thousands of proteins with PAS domains identifiable using the Pfam and SMART domain profiles.

Although PAS domains can be rapidly detected using the Pfam and SMART domain definitions, these profiles only identify the common PAS fold and fail to indicate the specific function of a given PAS domain. We chose to further characterize PAS domains because they represent a significant, extensive, multi-functional superfamily currently represented by a very general domain profile. Furthermore, there is substantial experimental data available to validate and corroborate this analysis including several known structures of PAS domains. In order to constrain the data for this study, analysis was restricted to PAS domains found in association with bacterial methyl-accepting chemotaxis proteins (MCPs). These proteins may be readily identified using the Pfam

and SMART domain profiles. The well-studied Aer transducer of *Escherichia coli* is a representative of this class of signal transduction proteins (Taylor, et al., 1999).

Here we demonstrate that chemotaxis PAS domains can be separated into functionally distinguishable subfamilies using protein sequence analysis. Based on extensive experimental data on the functional role of the PAS domain from the Aer protein, we were able to predict that all related members within this subfamily will bind FAD and sense changes in redox potential. The newly built HMM for this subfamily enables the automatic detection of many important microbial redox sensors in public databases. Additionally, this investigation revealed a distinct subfamily with substantial sequence conservation, which is predicted to bind a specific (although unknown) ligand and represents a good target for experimental studies. Using chemotaxis PAS domains as a case study, we have laid the groundwork for developing a more general approach to functionally resolving existing, multifunctional domains and increasing the accuracy and quality of annotation within public sequence databases.

Methods

HMM searches (Eddy, 1998) seeded with profile hidden Markov models from Pfam version 16.0 and SMART version 3.5, were carried out against the NR database (January 4, 2005). First, we identified 2021 MCP sequences in the NR database using the Pfam MCPsignal (accession PF00015) and SMART MA (accession SM00283) domain profiles. PAS domains in these MCPs were detected with the PAS and PAC domain profiles from both Pfam (accession PF00989 and PF00785) and SMART (accession SM00086 and SM00091). In many cases, only the PAS or the PAC element was detected. Since PAS domains always consist of both the PAS and PAC motifs, the “missing” motif

was delineated from pairwise alignments to other PAS domains from a PSI-BLAST (Altschul, et al., 1997) search with this region. If no significant hits were found, the PAS domain was excluded from the analysis. Current Pfam and SMART profiles fail to recognize all PAS domains. In order to address this problem, we used sensitive PSI-BLAST searches (inclusion threshold E = 0.01) initiated with the N-terminal regions of all 2021 MCPs retrieved from the NR-database. Additionally, we included several PAS domains that existed as a single protein yet were predicted to interact with MCPs based on genome neighborhood. These were confirmed experimentally (Hendrixson, et al., 2001). Partial or incomplete PAS domains and duplicate sequences were excluded from this analysis.

A multiple sequence alignment of PAS domains was constructed using the PCMA (Pei, et al., 2003) and ClustalW (Thompson, et al., 1994) programs and manually adjusted (Figure 6.1). The alignment was further tweaked using the VISSA visualization (Ulrich and Zhulin, 2005). Sequence conservation information and the visualized secondary structure of the PAS domains enabled us to produce a much better quality alignment than the initial, unedited alignment constructed by PCMA (Appendix C, Figures C.1 and C.2).

In order to depict the relationship between these PAS domains, we generated a neighbor-joining tree (Saitou and Nei, 1987) of the alignment based on sequence similarity (p-distance) using the MEGA3 software package (Kumar, et al., 2004). Profile HMMs were then built and calibrated for each major cluster within the neighbor-joining tree using the HMMER2 software package (http://hmmer.wustl.edu). To increase the sensitivity of each profile, the sequences were weighted based on the Krogh/Mitchison

maximum entropy algorithm (Krogh and Mitchison, 1995). Each specialized PAS HMM was then used in HMMER searches against the NR database to ensure that all members in the respective subfamily were retrieved as well as to identify any new PAS homologs. Strict and relaxed inclusion thresholds were determined and incorporated into the HMM for each PAS subfamily. The strict threshold was set as the bit score of the lowest scoring PAS domain of respective subfamily. The relaxed threshold was set as the bit score of the lowest scoring hit that was distinct from other PAS HMM searches. Searches with these cutoff thresholds revealed that both PAS_Aer and PAS_Che are completely specific and the difference in E-value between the last true positive and first true negative for each subfamily is given in Table 6.1. We evaluated the performance of these new PAS domain profiles and found that they significantly outperform all previous and current PAS profiles from Pfam and SMART (Table 6.2).

Table 6.1 E-values and scores of the last true positive and first true negative from searches of the PAS_Aer and PAS_Che profiles against the non-redundant database (4 January 2005). Scores are given in the parenthesis following the E-value.

Domain profiles	*Last true positive*	*First true negative*	*Difference*
PAS_Aer	2.771e-11 (50.24)	9.534e-11 (44.55)	6.736e-11 (5.684)
PAS_Che	1.707e-20 (86.78)	7.211e-12 (31.39)	7.211e-12 (55.39)

Table 6.2 Identification of PAS domains in methyl-accepting chemotaxis proteins by HMM domain profiles.

Domain profiles	*Number of complete PAS domains identified*	*Number of incomplete PAS domains identified*		*Total number of PAS domains identified*
		PAS motif only	*PAC motif only*	
Pfam 17.0 (PAS)	160	N/A	N/A	160
SMART 4.0 (PAS, PAC)	203	35	29	267
This study (PAS_Aer, PAS_Che, generic PAS)	274	N/A	N/A	274

Signal peptides and transmembrane regions with their topology were assigned using Phobius (Kall, et al., 2004). Consensus sequence patterns were generated using the CONSENSUS script by Nigel Brown (http://www.bork.embl-heidelberg.de/Alignment/consensus.html). Sequence logos based on the information content were created with the Berkeley weblogo server (Crooks, et al., 2004). A position in the MSA was considered strongly conserved if either the BLOSUM group consensus was greater than 85% or the information content was greater than three bits.

Results and Discussion

Searches with Pfam and SMART domain profiles yielded 240 complete and partial PAS domains associated with MCPs. PSI-BLAST searches revealed an additional twenty PAS domains and enabled the reconstruction of all partially detected PAS domains (e.g. PAC motif only) from domain searches. Finally, fourteen PAS domains comprising individual proteins, but predicted to interact with MCPs on the basis of genome neighborhood (encoded within the same operon), and experimental evidence

(Hendrixson, et al., 2001) were also included in this analysis. We constructed a multiple sequence alignment of these PAS domains and edited it using the VISSA protocol (Ulrich and Zhulin, 2005). The VISSA technique greatly facilitated the process of producing a high-quality alignment by revealing inconsistencies in the MSA such as misalignment of structural elements (Appendix C, Figures C.1 and C.2). Furthermore, VISSA corroborates these results by showing a high degree of agreement between the overall secondary structure of the alignment and the known structure of several PAS domains.

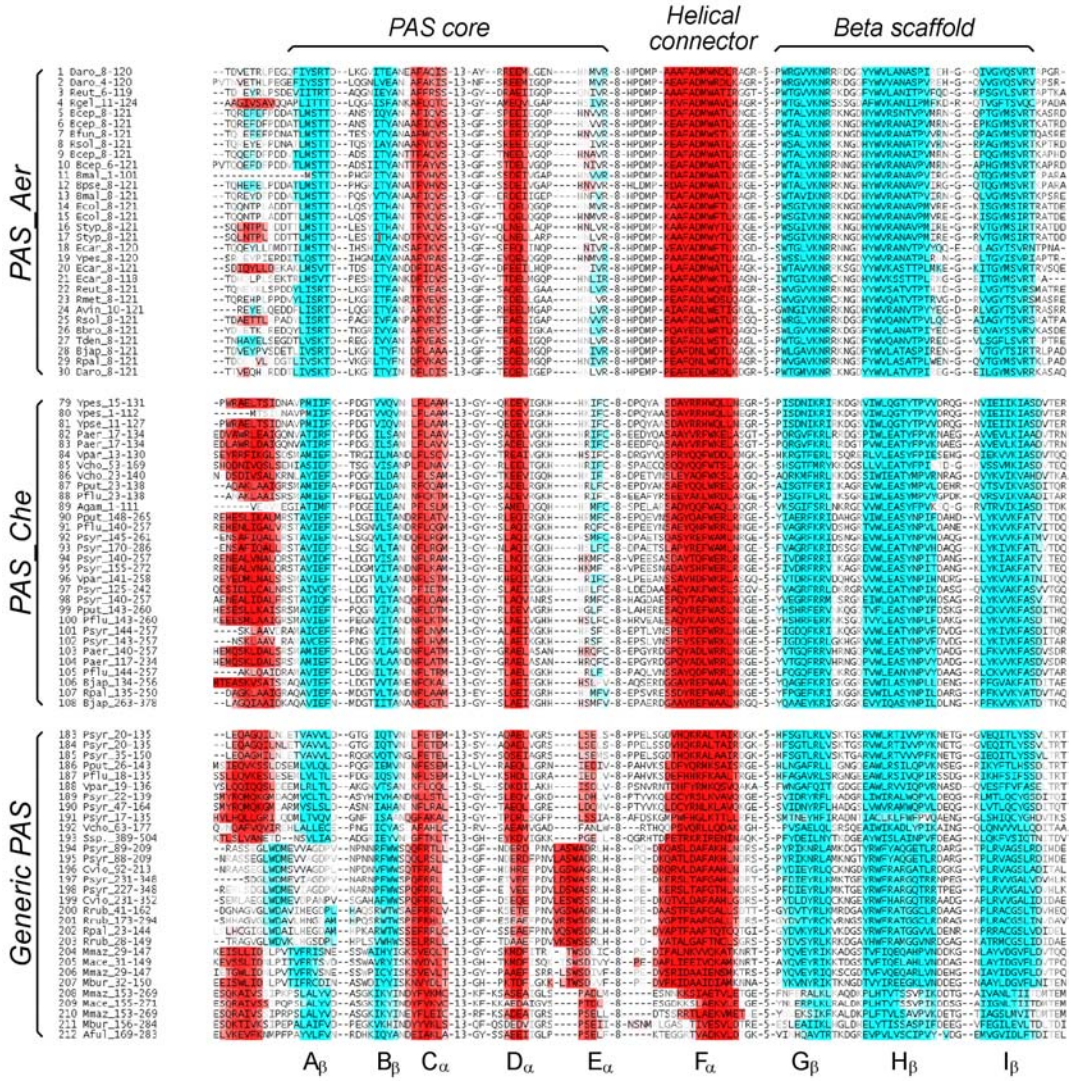


Figure 6.1 Multiple sequence alignment with the VISSA visualization of the three subfamilies of PAS domains found in methyl-accepting chemotaxis proteins. Thirty representative PAS domains from each subfamily are shown (for the full alignment of 274 sequences, see Appendix C, Figure C.2). The multiple alignment was constructed using the PCMA (Pei, et al., 2003) and ClustalW (Thompson, et al., 1994) programs, visualized using the VISSA protocol (Ulrich and Zhulin, 2005). Regions containing predicted alpha helices and beta strands have the background shaded in red and blue, respectively. The shading intensity is directly proportional to the confidence of a given prediction – a darker intensity representing a higher confidence. The PAS core, helical connector, and beta scaffold structural regions are delineated above the alignment. Structural elements, as defined by Gong *et al*. (1998), are listed below the alignment.

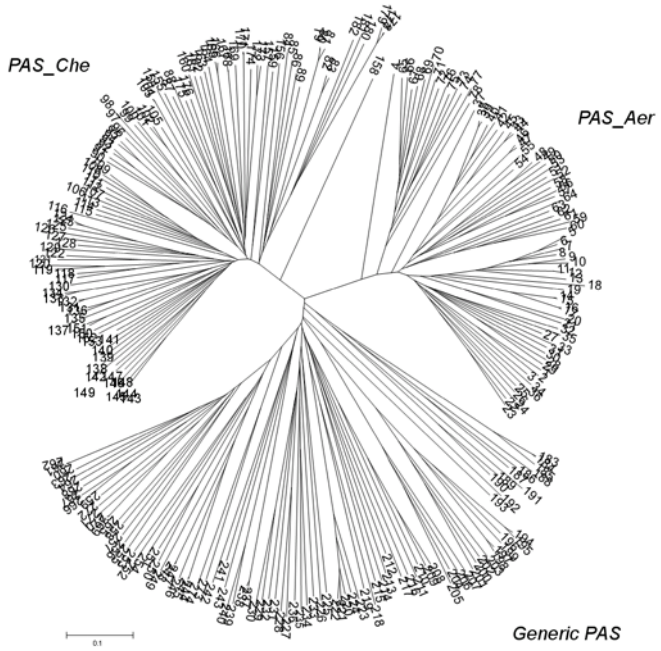


Figure 6.2 Neighbor-joining tree built from a multiple sequence alignment of 274 PAS domains showing three distinct subfamilies.

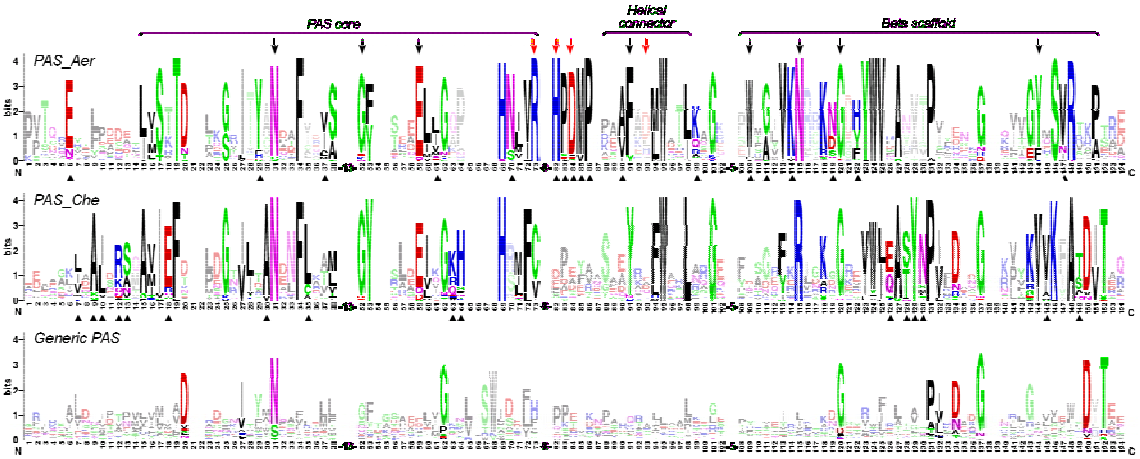


Figure 6.3 Sequence logos for three PAS subfamilies. Highlighted positions are strongly conserved (BLOSUM consensus at least 85% or

information content greater than three bits). Residues unique to a given subfamily are designated with a triangle (▲). Red arrows indicate residues critical for binding FAD by the PAS domain in *E. coli* Aer and black arrows indicate other residues that are critical for Aer function (Repik, et al., 2000).

A neighbor-joining tree built from this high quality alignment revealed two distinct, conserved clusters and one divergent set of PAS domains (Figure 6.2). We designated these three groups as PAS_Aer, PAS_Che, and generic PAS, respectively, based on sequence conservation and the domain architecture of proteins that contain these domains (see below). The secondary structure predictions for each PAS domain are largely consistent with known PAS structures. Six positions (I27, N31, G62, G120, P131, and G137) are strongly conserved across all three groups. Mutational studies indicate that N31 stabilizes the BC turn by forming hydrogen bonds to three backbone nitrogen atoms (Pellequer, et al., 1999; Pellequer, et al., 1998). P131 resides at the C-terminal end of H_β and in conjunction with G137, is responsible for the sharp HI loop of the beta scaffold. Twenty-four other residues are moderately conserved (information content greater than 2 bits) across the three groups (Figure 6.3).

The PAS_Aer subfamily consists of seventy-eight PAS domains including the PAS domain of the experimentally studied *E. coli* aerotaxis transducer, Aer. Fifty-seven amino acid positions are strongly conserved and eighteen of them are unique to this subfamily. All but one of these eighteen positions are conserved in the *E. coli* Aer protein. The most prominent feature that distinguishes this group is the conserved RHPDMP motif located within the EF loop (Figure 6.3). Experimental evidence indicates

that R73, H82, and D84 of this motif (the RHPDMP motif) are crucial for binding the FAD cofactor in Aer (Repik, et al., 2000). The predicted secondary structure of the PAS_Aer subfamily is largely consistent with the overall PAS fold except for the N-terminal cap, which lacks the well-defined alpha helix found in PAS_Che and generic PAS (Figure 6.1). Interestingly, a charged amino acid (predominantly E6) is uniquely conserved within this region indicating its possible functional role (Figure 6.3). Previous experimental studies (Repik, et al., 2000) identified twelve key residues in the PAS domain of Aer that were essential for producing an aerotactic response. Four of these residues (N31, G52, E59, and G120) are conserved throughout the entire alignment and loss of function by cysteine replacement was most likely related to disruption of the PAS fold. Four other residues (R73, H82, D84, and D93) are essential for FAD binding (see above). The remaining four residues (F91, W109, N115, F144) are located within secondary structure elements and show strong conservation only within the PAS_Aer subfamily (Figure 6.3).

Database searches with an HMM built for the PAS_Aer subfamily revealed that this version of the PAS domain is not unique to MCPs and can be found in both one- and two-component signal transduction systems (Ulrich, et al., 2005), such as di-guanylate cyclases and histidine kinases. The domain architecture of PAS_Aer containing MCPs is well conserved and typically includes a single PAS domain followed by one or two transmembrane regions, a HAMP domain, and the C-terminal MCP signaling domain. Occasionally, these (and only these) PAS domains exist as single proteins but are predicted to interact with the corresponding MCPs that contain a HAMP domain, but lack an N-terminal, sensing domain and reside adjacent to the PAS_Aer domains on the

chromosome. In addition to *E. coli* Aer, two other PAS_Aer-containing proteins have been experimentally studied. Both of them, the Aer homolog from *Pseudomonas putida* (Nichols and Harwood, 2000) and the CetA-CetB bipartite (PAS_Aer and the corresponding MCP) system from *Camplylobacter jejuni* (Hendrixson, et al., 2001) were shown to govern aerotaxis and related responses in these organisms.

The PAS_Che subfamily comprises 104 PAS domains that share strong sequence similarity. This subfamily contains fifty-eight strongly conserved residue positions and sixteen of them are unique (Figure 6.3). The predicted secondary structure of each of the PAS_Che domains agrees with the known PAS structures. The length of loops/turns in the PAS_Che subfamily is highly conserved (as it is in the PAS_Aer subfamily). Database searches with the PAS_Che profile confidently identified this type of PAS domains predominantly within methyl-accepting chemotaxis proteins (hence, the name PAS_Che); however, members of the PAS_Che domain subfamily can also be found in two other major classes of prokaryotic signal transduction, namely, histidine kinases and di-guanylate cyclases. MCPs containing PAS_Che domains have one to four copies of PAS_Che and occasionally one or two generic PAS domains. The prominent feature of this subfamily is that PAS_Che-containing MCPs lack any predicted transmembrane regions and therefore are likely to be exclusively cytoplasmic receptors. Only one PAS_Che-containing protein, the McpY of *Sinorhizobium meliloti* was experimentally studied and found not to be an aerotaxis/energy taxis transducer (Scharf, Schmitt, Zhulin and Taylor, unpublished data). Furthermore, although McpY contributes to the overall chemotactic response, the reconstituted McpY protein does not bind FAD (Meier,

Muschler and Scharf, unpublished data). Thus, the intracellular signal recognized by this MCP remains unknown.

The third group of PAS domains found in MCPs consists of ninety-two members and both the alignment and neighbor-joining tree reveal a highly divergent set of sequences. The predicted secondary structure still follows that of a typical PAS fold, although several sequences contain an extension of the C_α helix and appear to be missing the E_α helix. Thirty-five PAS domains in this subset have a reduced helical connector and a shorter G_β strand. Only eleven residues are conserved in generic PAS domains and none of these positions are unique to this subset. Such remarkable lack of conservation is in striking contrast to the PAS_Aer and PAS_Che subfamilies and suggests that no specific function is associated with this group (hence the name, "generic PAS"). Searches with the profile HMM built for this group of sequences recognize PAS domains in numerous proteins representing all major classes of prokaryotic signal transduction. In MCPs, generic PAS domains occur in a single protein in up to five copies and can co-occur with either PAS_Aer or PAS_Che domains.

No MCPs that contain generic PAS domains have been studied experimentally and elucidation of their function, as well as the function of many other PAS domain-containing signal transduction proteins, necessitates high-quality, high-throughput sequence analysis of the entire PAS superfamily.

In summary, the high-quality, microbial signal transduction data derived by these analytical techniques offer promising new perspectives and expectations from the available, underlying genetic information.

Acknowledgements

We extend our thanks to our co-authors: William Black, Qinhong Ma, and Barry Taylor. We also thank Meier, Muschler, Scharf, Schmitt, Zhulin, and Taylor for their supportive role in providing unpublished results.

CHAPTER 7

CONCLUSION

In conclusion, this research significantly contributes to the understanding of microbial signal transduction. A bioinformatics platform and the Microbial Signal Transduction database (MiST) were developed that enabled the comprehensive, integrated analysis of microbial signal transduction at the domain level. This platform was used successfully for the high-throughput identification and classification of signal transduction systems in more than 300 bacterial and archaeal organisms. Using this infrastructure, we performed a comprehensive review of signal transduction systems and found that contrary to the current view, the majority of signal transduction systems consist of one-component systems – a single protein containing input and output domains but lacking phosphotransfer domains typical of two-component systems. Two-component systems utilize a subset of the input and output domains found in one-component systems, indicating a reduced repertoire of sensory and response modules. Furthermore, one-component systems display extraordinary combinatorial domain diversity. The dominance of one-component systems in prokaryotic signal transduction is a paradigm-shifting discovery rendering a more accurate depiction of the scope and complexity of bacterial signal transduction. A novel source of information regarding signal transduction systems, the MiST database has been effectively used for annotating novel genomes and performing indepth characterization of sensory domains. Altogether, this systematic, high-throughput delineation of microbial signal transduction is another step forward in our understanding of the genomic basis of life.

APPENDIX A

SUPPLEMENTARY INFORMATION FOR CHAPTER 4

Table A.1 Domains and domain categories used to identify signal transduction systems.

Input domains

Small-molecule binding	*Enzymatic*	*Cofactor binding*	*Unknown function*	*Protein-protein interactions*
ACT	aminotran_1_2	BLUF	CHASE2	CBS
Ada_Zn_binding	Archaeal_ATPase	fer4	CHASE3	TPR
AlkA_N	citrate_synt	FeS	CHASE4	
AraC_binding	Cyanate_lyase	Hemerythrin	MASE1	
Autoind_binding	EPSP_synthase	HhH-GPD	MASE2	
Cache	FmddA_AmdA	NIR_SIR	MHYT	
CHASE	GATase_2	NIR_SIR_ferr	TrkA-C	
cNMP_binding	Glucokinase	Nitro_FeMo-Co		
Diacid_rec	Glycos_trans_3N	phytochrome		
Fe_dep_repr_C	Glyoxalase			
FeoA	HEAT_PBS			
FHA	HEM4			
GAF	Nitroreductase			
HMA	NTP_transf_2			
LysR_substrate	NUDIX			
PAS	PALP			
Peripla_BP_2	Peptidase_M37			
Peripla_BP_like	peroxidase			
SBP_bac_3	pfkB			
SIS	Pribosyltran			
STAS	PTS_EIIC			
TOBE	PTS-HPr			
V4R	pyr_redox			
NIT*	Rhodanese			
	SKI			

Output domains

DNA-binding	*Di-guanylate cyclase*	*RNA-binding*	*Phosphatase*	*Protein kinase*	*Hydrolase*	*Other*
Arc	EAL	ANTAR	PP2C_SIG**	pkinase	HD	guanylate_cyc
Arg_repressor	GGDEF	CsrA				LytR_cpsA_psr
ASNC_trans_reg						Rrf2
crp						RseA_N
CtsR						
deoR						
Fe_dep_repress						
GerE						
gntR						

Table A.1 (continued)

HTH_1
HTH_3
HTH_4
HTH_5
HTH_6
HTH_7
HTH_8
HTH_9
HTH_10
HTH_AraC
IclR
AlcI
LytTR
MarR
MerR
PadR
ROS_MUCR
TetR
trans_reg_C

Transmitter / receiver domains

Histidine kinases	*Response regulator*	*Other*
HATPase_c	response_reg	HAMP
HisKA		MCPsignal
		Sigma54_activat

All domain nomenclature and assignments are according to Pfam (ref)
* Domain is from (ref)
** Domain is from the SMART database (ref)

Table A.2 Distribution of two-component and one-component systems in prokaryotic genomes.

	Genome size (Mb)	Two-component systems		One-component systems
		Histidine kinases	Response regulators	
Archaea				
Crenarchaeota				
Aeropyrum pernix	1.7	0	0	14
Pyrobaculum aerophilum	2.2	0	0	25
Sulfolobus solfataricus	3	0	0	49
Sulfolobus tokodaii	2.7	0	0	51
Euryachaeota				
Archaeoglobus fulgidus	2.2	14	11	49

Table A.2 (continued)

Methanosarcina acetivorans	5.8	53	15	66
Methanosarcina mazei	4.1	34	14	50
Halobacterium sp.	2.6	13	5	56
Methanobacterium thermoautotrophicum	1.8	16	7	24
Methanococcus jannaschii	1.7	1	0	29
Methanosarcina barkeri	5.1	23	10	47
Methanopyrus kandleri	1.7	0	0	17
Pyrococcus abyssi	1.8	1	2	32
Pyrococcus furiosus	1.9	0	0	30
Pyrococcus horikoshii	1.7	1	2	30
Thermoplasma acidophilum	1.6	0	0	22
Thermoplasma volcanium	1.6	0	0	21
Ferroplasma acidarmanus	1.9	0	0	32
Bacteria				
Aquificae				
Aquifex aeolicus	1.6	3	5	34
Thermotogae				
Thermotoga maritima	1.9	10	11	57
Planctomycetes				
Pirellula sp.	7.1	45	56	137
Cyanobacteria				
Synechococcus sp.	2.4	6	9	16
Synechocystis PCC6803	3.6	42	40	68
Thermosynechococcus elongatus	2.6	19	23	40
Nostoc sp.	7.2	137	80	148
Nostoc punctiforme	9	164	100	146
Trichodesmium erythraeum	7.7	35	28	69
Prochlorococcus marinus	1.8	4	6	9
Deinococcus				
Deinococcus radiodurans	3.3	20	24	102
Actinobacteria				
Bifidobacterium longum	2.3	10	9	67
Corynebacterium efficiens	3.1	13	13	91
Corynebacterium glutamicum	3.3	13	13	111
Mycobacterium bovis	4.3	15	12	162
Mycobacterium leprae	3.3	5	5	38
Mycobacterium tuberculosis	4.4	15	12	159
Streptomyces avermitilis	9	132	73	519
Streptomyces coelicolor	9.1	143	85	627
Tropheryma whipplei	0.9	2	2	7
Thermobifida fusca	3.6	40	26	159
Firmicutes				
Bacillus anthracis Ames	5.2	49	47	230

Table A.2 (continued)

Bacillus cereus	5.4	55	46	224
Bacillus halodurans	4.2	46	48	174
Bacillus subtilis	4.2	39	35	187
Listeria innocua	3.1	18	18	150
Listeria monocytogenes	2.9	17	17	147
Oceanobacillus iheyensis	3.6	22	21	145
Staphylococcus aureus	2.9	18	17	72
Staphylococcus epidermidis	2.5	17	16	50
Clostridium acetobutylicum	4.1	38	43	174
Clostridium perfringens	3.1	28	20	95
Clostridium tetani	2.8	31	28	88
Thermoanaerobacter tengcongensis	2.7	20	21	79
Clostridium thermocellum	3.7	32	32	75
Desulfitobacterium hafniense	7	67	81	273
Enterococcus faecalis	3.2	15	18	138
Lactobacillus plantarum	3.3	11	15	189
Lactococcus lactis	2.4	8	7	96
Streptococcus agalactiae	2.2	18	21	67
Streptococcus mutans	2	13	14	90
Streptococcus pneumoniae	2.2	13	14	61
Streptococcus pyogenes	1.9	10	12	67
Enterococcus faecium	3	13	16	92
Lactobacillus gasseri	2	5	5	65
Leuconostoc mesenteroides	2.1	7	7	77
Oenococcus oeni	1.9	6	6	63
Mycoplasma gallisepticum	1	0	0	2
Mycoplasma genitalium	0.6	0	0	4
Mycoplasma penetrans	1.4	0	0	9
Mycoplasma pneumoniae	0.8	0	0	3
Mycoplasma pulmonis	1	0	0	1
Ureaplasma urealyticum	0.8	0	0	3
Fusobacteria				
Fusobacterium nucleatum	2.2	8	7	37
Fusobacterium nucleatum subsp. vincentii	2.1	6	4	31
Chlamydiae				
Chlamydia muridarum	1.1	2	1	6
Chlamydia trachomatis	1	2	1	6
Chlamydophila caviae	1.2	2	1	8
Chlamydophila pneumoniae	1.2	2	1	7
Spirochaetes				
Borrelia burgdorferi	1.5	5	6	3
Leptospira interrogans	4.7	48	32	101
Treponema pallidum	1.1	3	3	13
Bacteroidetes				
Bacteroides thetaiotaomicron	6.3	76	30	98

Table A.2 (continued)

Porphyromonas gingivalis	2.3	6	6	24
Cytophaga hutchinsonii	4.4	56	32	60
Chlorobi				
Chlorobium tepidum	2.2	9	4	17
Chloroflexi				
Chloroflexus aurantiacus	4.9	70	66	106
Proteobacteria - epsilon				
Campylobacter jejuni	1.6	7	11	9
Helicobacter hepaticus	1.8	5	10	10
Helicobacter pylori	1.7	3	9	5
Wolinella succinogenes	2.1	38	43	53
Proteobacteria - delta				
Desulfovibrio desulfuricans	3.9	51	66	112
Geobacter metallireducens	4.2	88	74	101
Proteobacteria - magnetotactic				
Magnetococcus sp.	4.7	95	73	108
Proteobacteria - alpha				
Magnetospirillum magnetotacticum	9.2	126	144	246
Novosphingobium aromaticivorans	4.4	27	33	156
Rhodobacter sphaeroides	4.6	41	48	170
Rhodopseudomonas palustris	5.5	55	55	227
Rhodospirillum rubrum	4.7	53	55	186
Agrobacterium tumefaciens	5.7	46	55	353
Bradyrhizobium japonicum	9.1	85	93	446
Brucella melitensis	3.3	20	22	133
Brucella suis	3.3	19	21	131
Caulobacter crescentus	4	53	46	151
Mesorhizobium loti	7.6	52	56	461
Rickettsia conorii	1.3	4	4	8
Rickettsia prowazekii	1.1	4	4	4
Sinorhizobium meliloti	6.7	40	56	390
Rickettsia sibirica	1.3	4	4	8
Proteobacteria - beta				
Burkholderia fungorum	9.6	70	73	545
Ralstonia metallidurans	6.8	61	84	377
Bordetella bronchiseptica	5.3	26	31	382
Bordetella parapertussis	4.8	22	28	312
Bordetella pertussis	4.1	19	21	221
Chromobacterium violaceum	4.8	47	64	233
Neisseria meningitidis	2.3	4	4	30
Nitrosomonas europaea	2.8	15	18	51
Ralstonia solanacearum	5.8	45	61	277
Proteobacteria - gamma				

Table A.2 (continued)

Azotobacter vinelandii	5.4	40	37	225
Microbulbifer degradans	5.4	62	72	156
Blochmannia floridanus	0.7	0	0	5
Buchnera aphidicola	0.6	0	0	1
Buchnera sp.	0.7	0	0	2
Coxiella burnetii	2	6	8	18
Escherichia coli	4.6	30	32	230
Haemophilus ducreyi	1.7	1	3	24
Haemophilus influenzae	1.8	3	6	42
Pasteurella multocida	2.3	9	9	41
Pseudomonas aeruginosa	6.3	64	71	394
Pseudomonas putida	6.2	68	73	350
Pseudomonas syringae	6.4	67	69	283
Salmonella typhi	5.1	30	34	236
Salmonella typhimurium	5	32	35	245
Shewanella oneidensis	5.1	46	57	204
Shigella flexneri	4.6	25	29	166
Vibrio cholerae	4	43	49	195
Vibrio parahaemolyticus	5.2	50	55	287
Vibrio vulnificus	5.1	52	58	283
Wigglesworthia brevipalpis	0.7	0	1	3
Xanthomonas campestris	5.1	55	54	163
Xanthomonas citri	5.2	61	59	167
Xylella fastidiosa	2.7	13	17	40
Xylella fastidiosa Temecula1	2.5	13	18	28
Yersinia pestis	4.8	21	27	198
Actinobacillus pleuropneumoniae	2.3	3	2	28
Haemophilus somnus	2.2	2	2	34
Pseudomonas fluorescens	6.5	75	84	362

Table A.3 Genomic distribution of input and output domains in archaea and bacteria.

	Two-component systems		One-component systems	
Input domains*	Archaea	Bacteria	Archaea	Bacteria
LysR_substrate	0	0	3	2542
PAS	104	485	1	470
GAF	21	284	0	317
Peripla_BP_like	0	3	0	537
AraC_binding	0	0	0	212
cNMP_binding	0	7	0	177
Cache	4	109	0	5
SIS	0	0	1	117

Table A.3 (continued)

aminotran_1_2	0	0	0	76
TPR	0	7	1	64
CBS	0	10	11	42
Fe_dep_repr_C	0	0	18	43
MHYT	0	3	0	54
SBP_bac_3	0	37	0	20
CHASE3	0	47	0	6
ACT	0	0	0	52
CHASE2	0	12	0	37
Ada_Zn_binding	0	0	0	46
FHA	0	2	0	39
NTP_transf_2	0	0	0	38
NIT	0	32	0	5
CHASE	0	16	0	21
Autoind_bind	0	0	0	35
CHASE4	10	0	0	20
TOBE	0	0	2	28

	Two-component systems		One-component systems	
Output domains*	Archaea	Bacteria	Archaea	Bacteria
HTH_1	0	0	5	2673
tetR	0	0	33	1634
gntR	0	0	6	1383
trans_reg_C	0	1202	0	176
HTH_AraC	0	85	0	1287
GGDEF	0	132	0	1221
HTH_3	0	0	104	1237
GerE	0	618	2	376
MarR	0	0	36	800
EAL	0	71	0	727
HD	0	53	79	570
HTH_8	0	263	0	334
HTH_5	0	0	98	461
lacI	0	0	0	537
merR	0	0	4	493
ASNC_trans_reg	0	0	93	399
pkinase	0	3	13	466
IclR	0	0	1	379
deoR	0	0	0	285
PP2C_SIG	0	58	1	200
PadR	0	0	39	187
Rrf2	0	0	0	222
guanylate_cyc	0	12	0	185
crp	0	1	1	194
LytTR	0	109	0	51

* Twenty-five most abundant domains are listed

APPENDIX B

SUPPLEMENTARY INFORMATION FOR CHAPTER 5

Figure B.1 Unedited seed 4HB_MCP alignment with VISSA visualization.

Figure B.1 continued

Figure B.1 continued

Figure B.1 continued

Figure B.1 continued

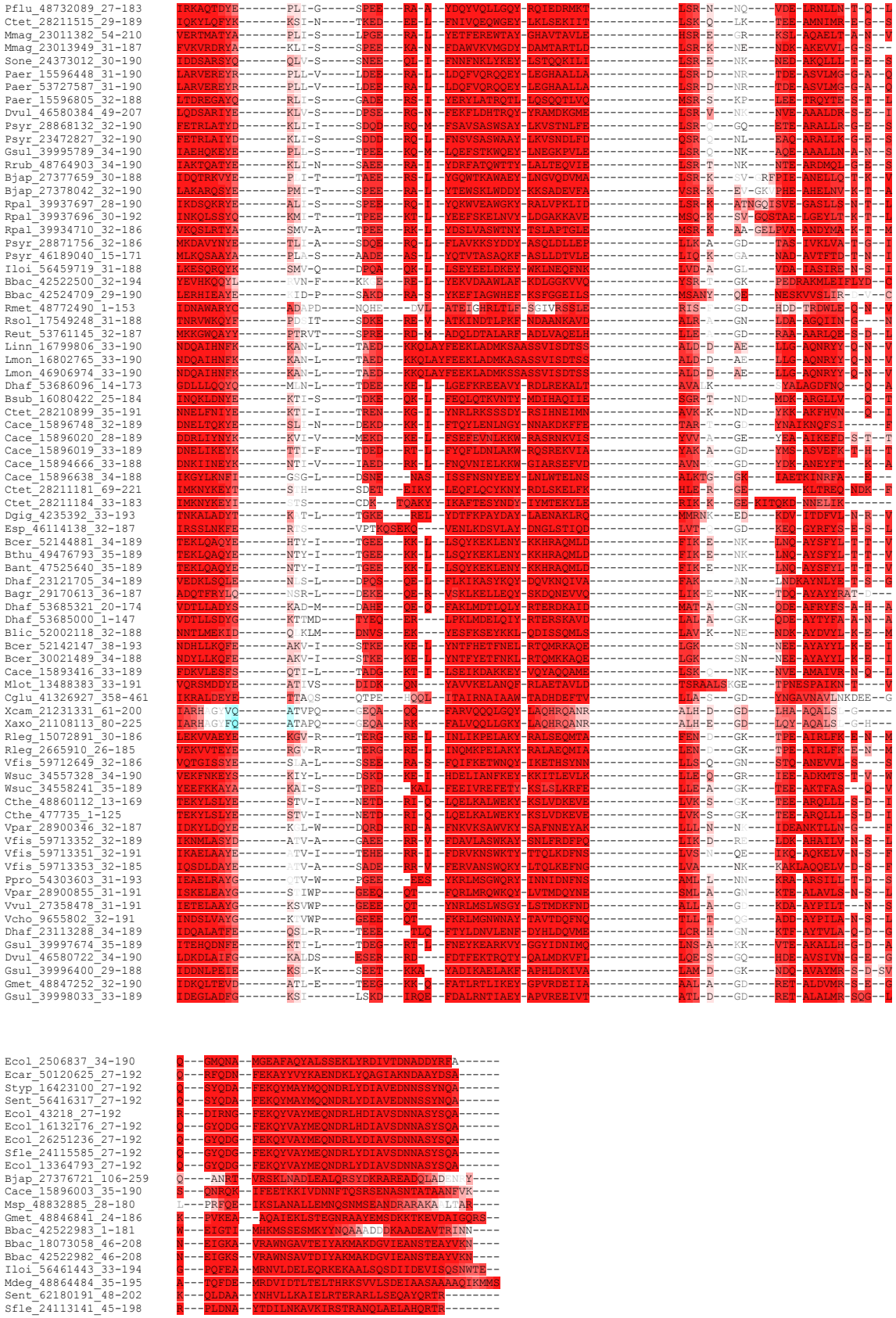


Figure B.1 continued

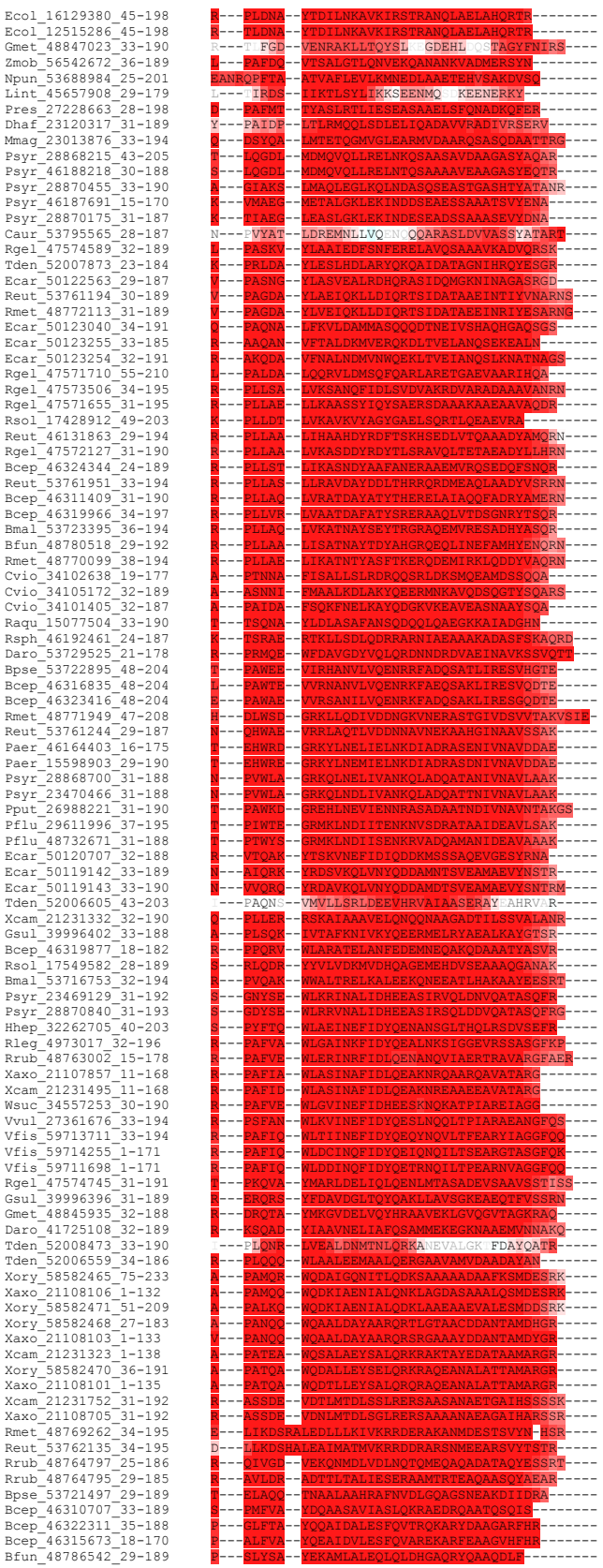


Figure B.1 continued

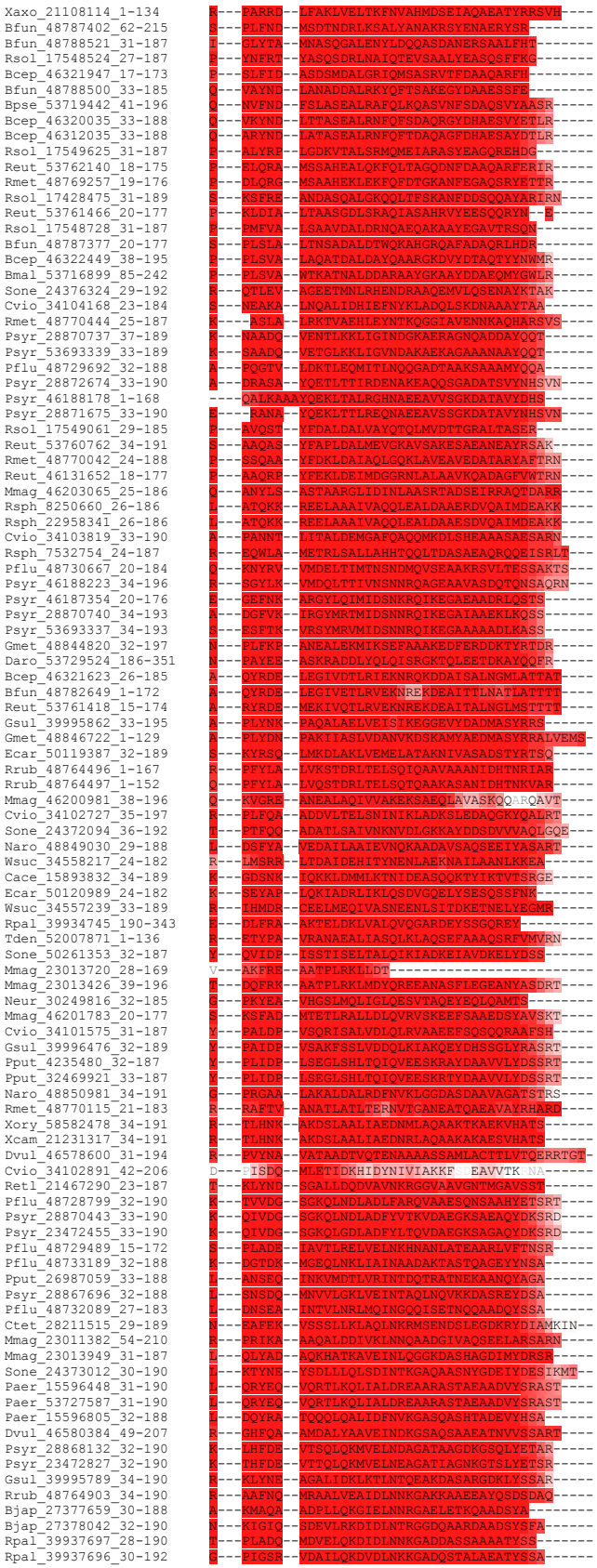
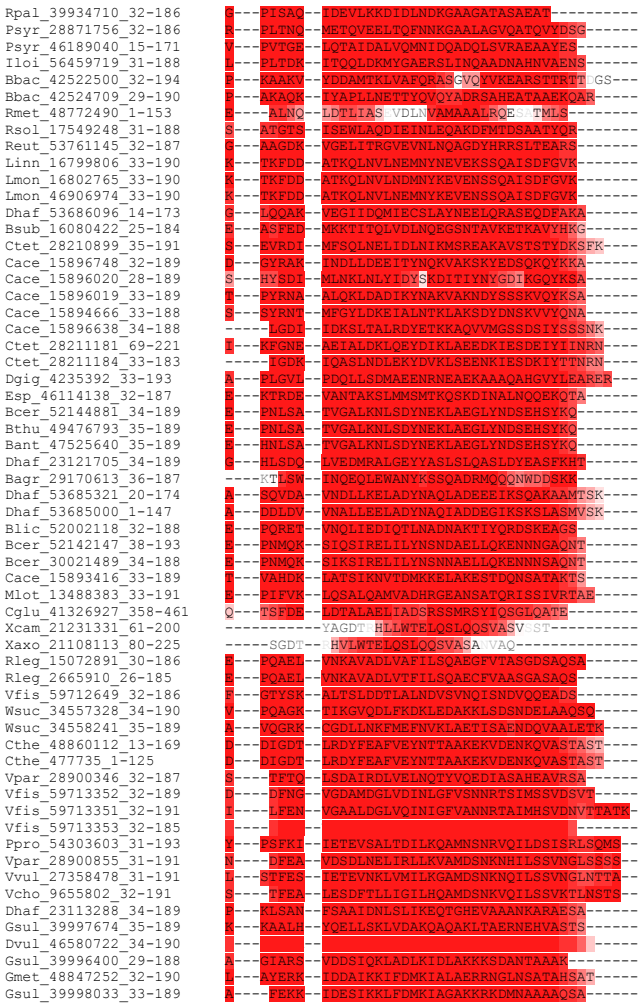


Figure B.1 continued



Bagr, *Brevibacillus agri*; Bant, *Bacillus anthracis*; Bbac, *Bdellovibrio bacteriovorus*; Bcep, *Burkholderia cepacia*; Bcer, *Bacillus cereus*; Bfun, *Burkholderia fungorum*; Bjap, *Bradyrhizobium japonicum*; Blic, *Bacillus licheniformis*; Bmal, *Burkholderia mallei*; Bpse, *Burkholderia pseudomallei*; Bsub, *Bacillus subtilis*; Bthu, *Bacillus thuringiensis*; Cace, *Clostridium acetobutylicum*; Caur, *Chloroflexus aurantiacus*; Cglu, *Corynebacterium glutamicum*; Ctet, *Clostridium tetani*; Cthe, *Clostridium thermocellum*; Cvio, *Chromobacterium violaceum*; Daro, *Dechloromonas aromatica*; Dgig, *Desulfovibrio gigas*; Dhaf, *Desulfitobacterium hafniense*; Dvul, *Desulfovibrio vulgaris*; Ecar, *Erwinia carotovora*; Ecol, *Escherichia coli*; Esp, *Exiguobacterium sp*; Gmet, *Geobacter metallireducens*; Gsul, *Geobacter sulfurreducens*; Hhep, *Helicobacter hepaticus*; Iloi, *Idiomarina loihiensis*; Linn, *Listeria innocua*; Lint, *Leptospira interrogans*; Lmon, *Listeria monocytogenes*; Mdeg, *Microbulbifer degradans*; Mlot, *Mesorhizobium loti*; Mmag, *Magnetospirillum magnetotacticum*; Msp, *Magnetococcus sp*; Naro, *Novosphingobium aromaticivorans*; Neur, *Nitrosomonas europaea*; Npun, *Nostoc punctiforme*; Paer, *Pseudomonas aeruginosa*; Pflu, *Pseudomonas fluorescens*; Ppro, *Photobacterium profundum*; Pput, *Pseudomonas putida*; Pres, *Pseudomonas resinovorans*; Psyr, *Pseudomonas syringae*; Raqu, *Rahnella aquatilis*; Retl, *Rhizobium etli*; Reut, *Ralstonia eutropha*; Rgel, *Rubrivivax gelatinosus*; Rleg, *Rhizobium leguminosarum*; Rmet, *Ralstonia metallidurans*; Rpal, *Rhodopseudomonas palustris*; Rrub, *Rhodospirillum rubrum*; Rsol, *Ralstonia solanacearum*; Rsph, *Rhodobacter sphaeroides*; Sent, *Salmonella enterica*; Sfle, *Shigella flexneri*; Sone, *Shewanella oneidensis*; Styp, *Salmonella typhimurium*; Tden, *Thiobacillus denitrificans*; Vcho, *Vibrio cholerae*; Vfis, *Vibrio fischeri*; Vpar, *Vibrio parahaemolyticus*; Vvul, *Vibrio vulnificus*; Wsuc, *Wolinella succinogenes*; Xaxo, *Xanthomonas axonopodis*; Xcam, *Xanthomonas campestris*; Xory, *Xanthomonas oryzae*; Zmob, *Zymomonas mobilis*

Figure B.1 continued

Figure B.2 Edited seed 4HB_MCP alignment with VISSA visualization.

Figure B.2 continued

Figure B.2 continued

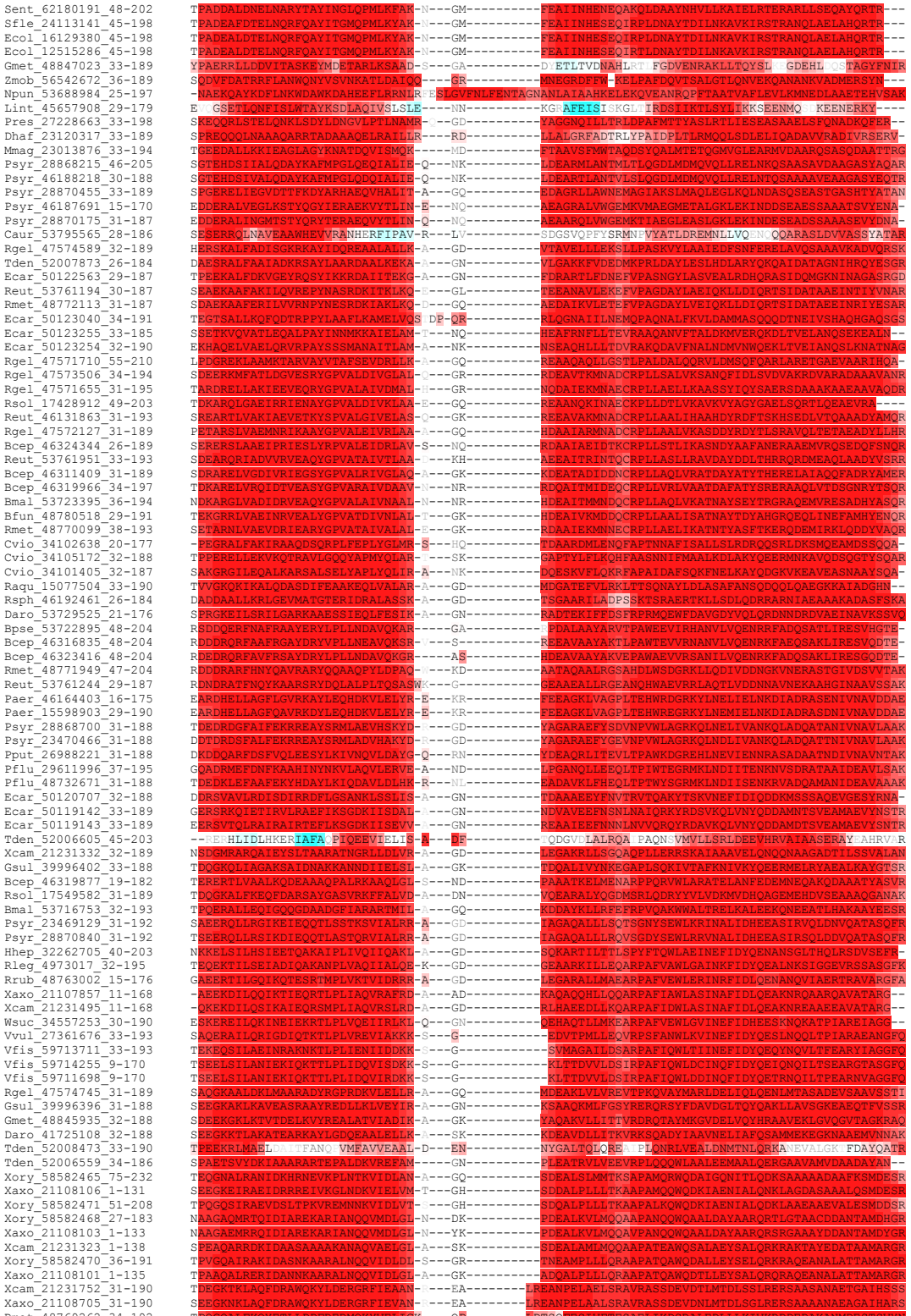
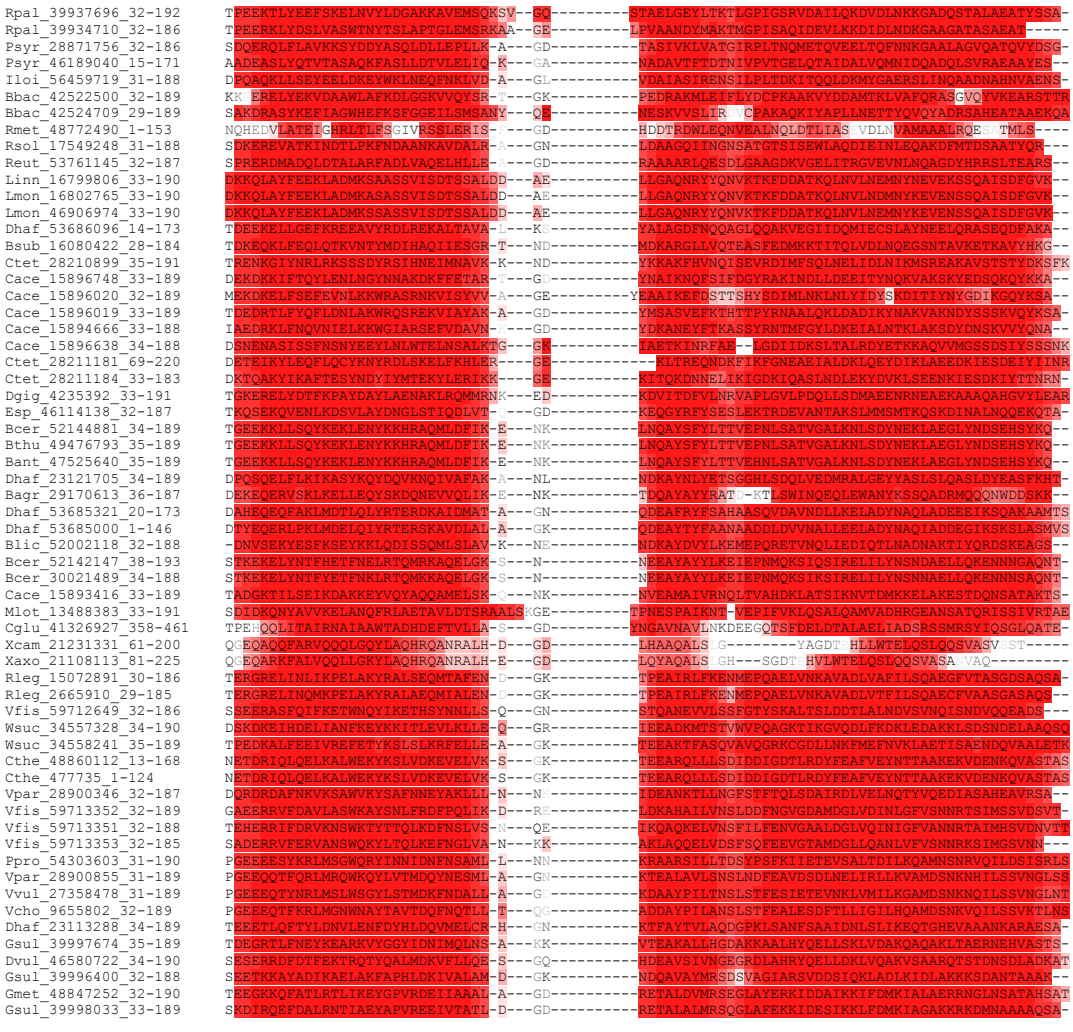


Figure B.2 continued

Figure B.2 continued



Bagr, *Brevibacillus agri*; Bant, *Bacillus anthracis*; Bbac, *Bdellovibrio bacteriovorus*; Bcep, *Burkholderia cepacia*; Bcer, *Bacillus cereus*; Bfun, *Burkholderia fungorum*; Bjap, *Bradyrhizobium japonicum*; Blic, *Bacillus licheniformis*; Bmal, *Burkholderia mallei*; Bpse, *Burkholderia pseudomallei*; Bsub, *Bacillus subtilis*; Bthu, *Bacillus thuringiensis*; Cace, *Clostridium acetobutylicum*; Caur, *Chloroflexus aurantiacus*; Cglu, *Corynebacterium glutamicum*; Ctet, *Clostridium tetani*; Cthe, *Clostridium thermocellum*; Cvio, *Chromobacterium violaceum*; Daro, *Dechloromonas aromatica*; Dgig, *Desulfovibrio gigas*; Dhaf, *Desulfitobacterium hafniense*; Dvul, *Desulfovibrio vulgaris*; Ecar, *Erwinia carotovora*; Ecol, *Escherichia coli*; Esp, *Exiguobacterium sp*; Gmet, *Geobacter metallireducens*; Gsul, *Geobacter sulfurreducens*; Hhep, *Helicobacter hepaticus*; Iloi, *Idiomarina loihiensis*; Linn, *Listeria innocua*; Lint, *Leptospira interrogans*; Lmon, *Listeria monocytogenes*; Mdeg, *Microbulbifer degradans*; Mlot, *Mesorhizobium loti*; Mmag, *Magnetospirillum magnetotacticum*; Msp, *Magnetococcus sp*; Naro, *Novosphingobium aromaticivorans*; Neur, *Nitrosomonas europaea*; Npun, *Nostoc punctiforme*; Paer, *Pseudomonas aeruginosa*; Pflu, *Pseudomonas fluorescens*; Ppro, *Photobacterium profundum*; Pput, *Pseudomonas putida*; Pres, *Pseudomonas resinovorans*; Psyr, *Pseudomonas syringae*; Raqu, *Rahnella aquatilis*; Retl, *Rhizobium etli*; Reut, *Ralstonia eutropha*; Rgel, *Rubrivivax gelatinosus*; Rleg, *Rhizobium leguminosarum*; Rmet, *Ralstonia metallidurans*; Rpal, *Rhodopseudomonas palustris*; Rrub, *Rhodospirillum rubrum*; Rsol, *Ralstonia solanacearum*; Rsph, *Rhodobacter sphaeroides*; Sent, *Salmonella enterica*; Sfle, *Shigella flexneri*; Sone, *Shewanella oneidensis*; Styp, *Salmonella typhimurium*; Tden, *Thiobacillus denitrificans*; Vcho, *Vibrio cholerae*; Vfis, *Vibrio fischeri*; Vpar, *Vibrio parahaemolyticus*; Vvul, *Vibrio vulnificus*; Wsuc, *Wolinella succinogenes*; Xaxo, *Xanthomonas axonopodis*; Xcam, *Xanthomonas campestris*; Xory, *Xanthomonas oryzae*; Zmob, *Zymomonas mobilis*

Figure B.2 continued

Figure B.3 Complete 4HB_MCP alignment with VISSA visualization.

Figure B.3 continued

Figure B.3 continued

Figure B.3 continued

Figure B.3 continued

Figure B.3 continued

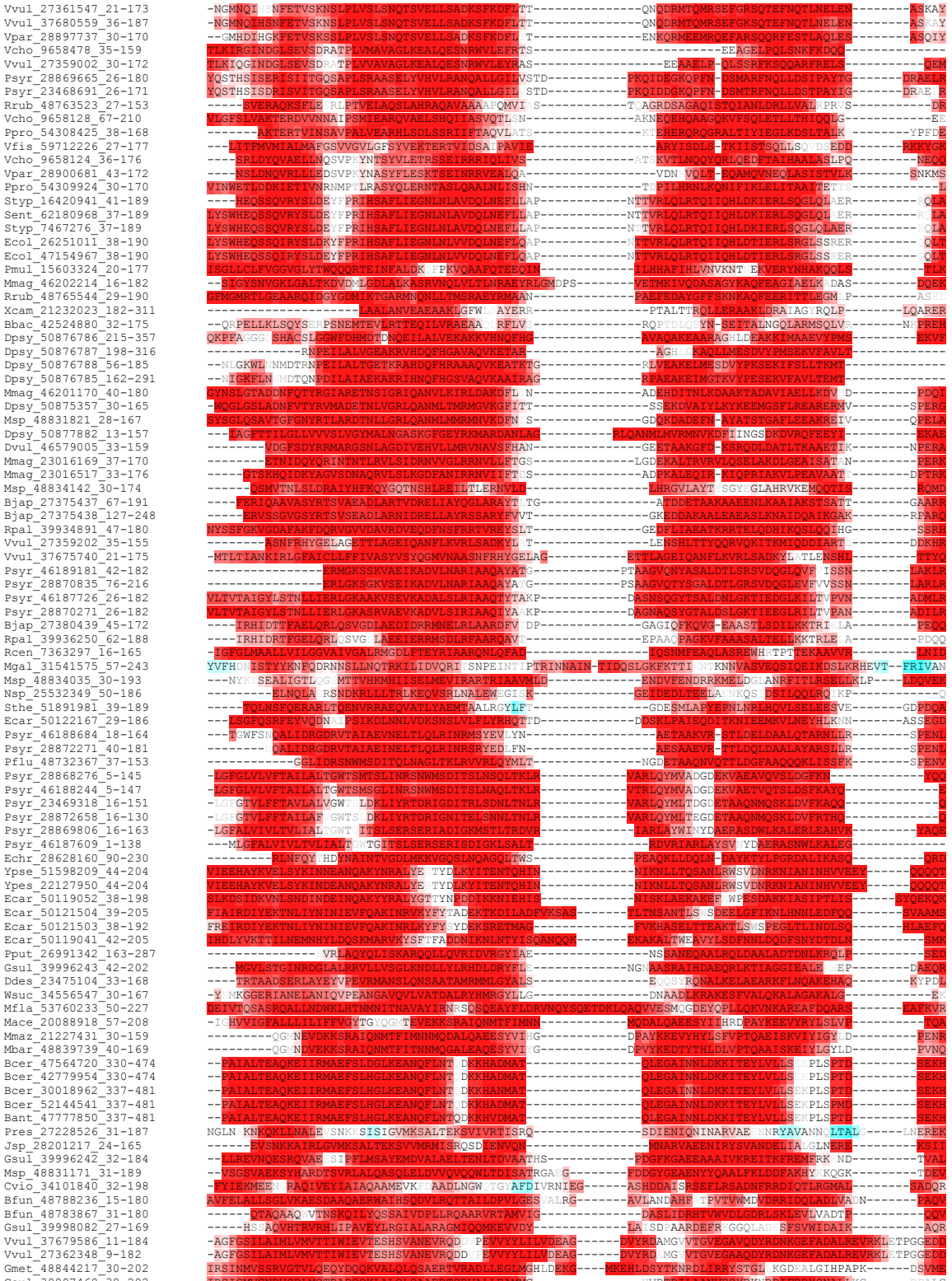


Figure B.3 continued

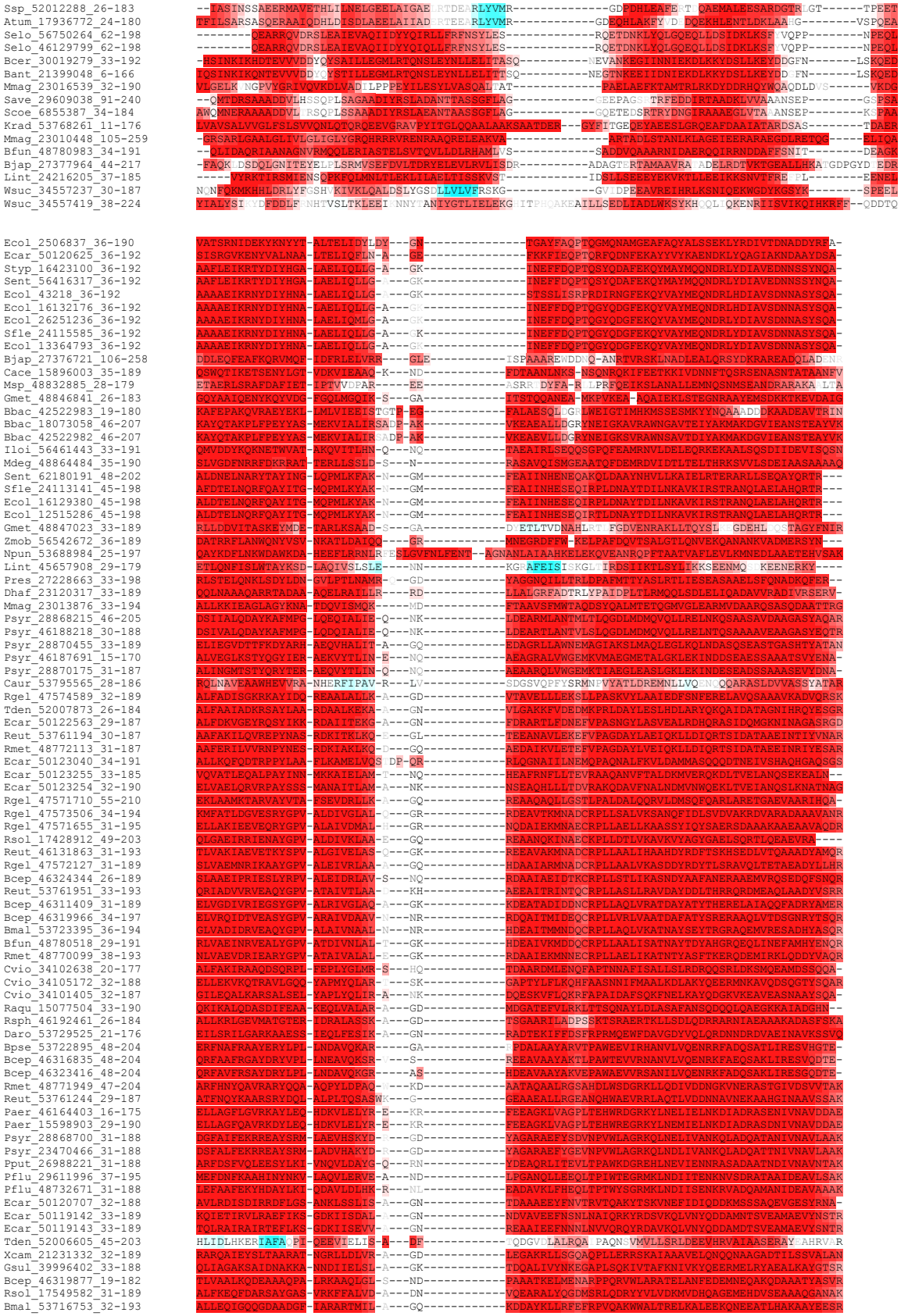


Figure B.3 continued

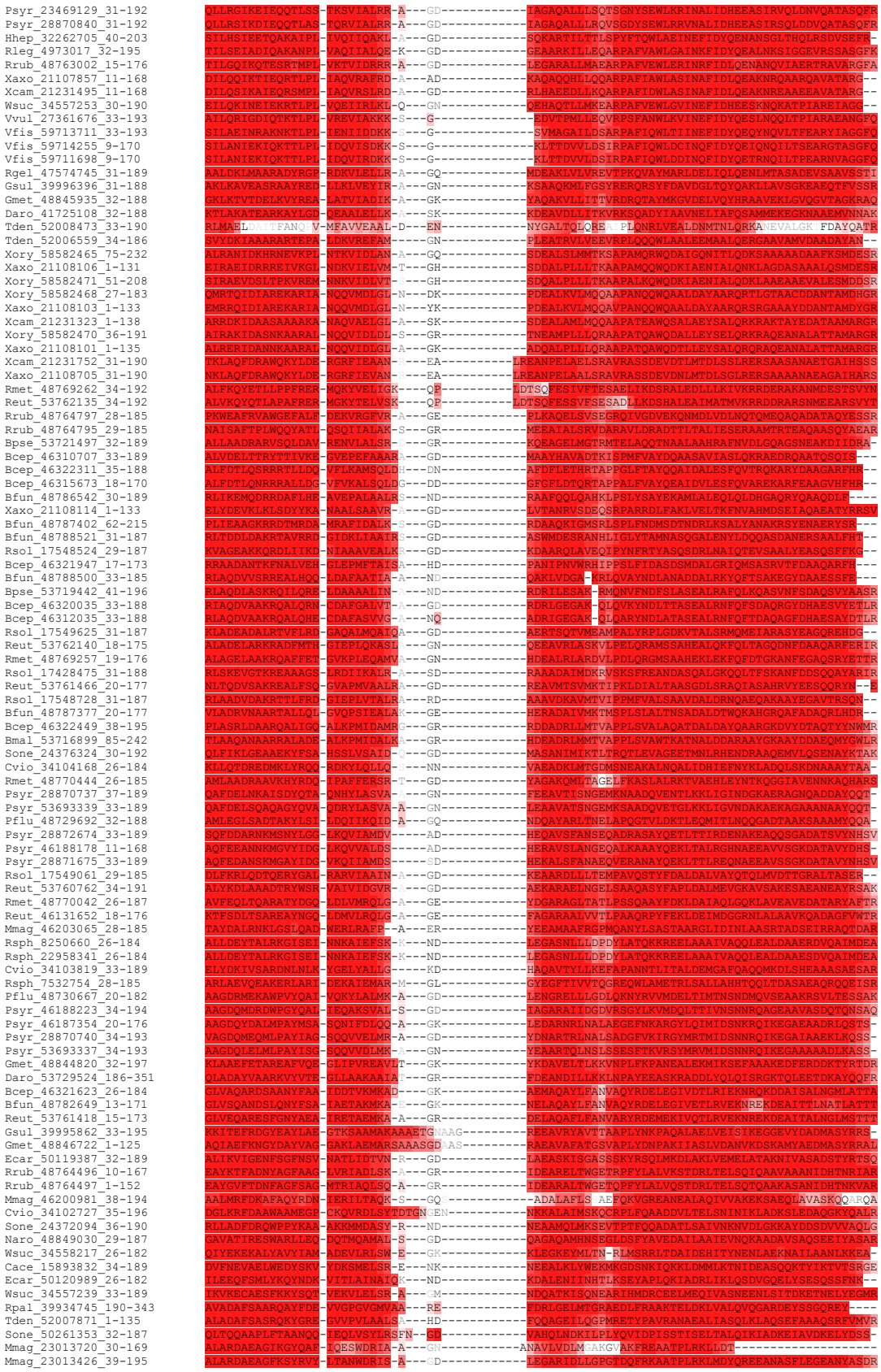


Figure B.3 continued

Figure B.3 continued

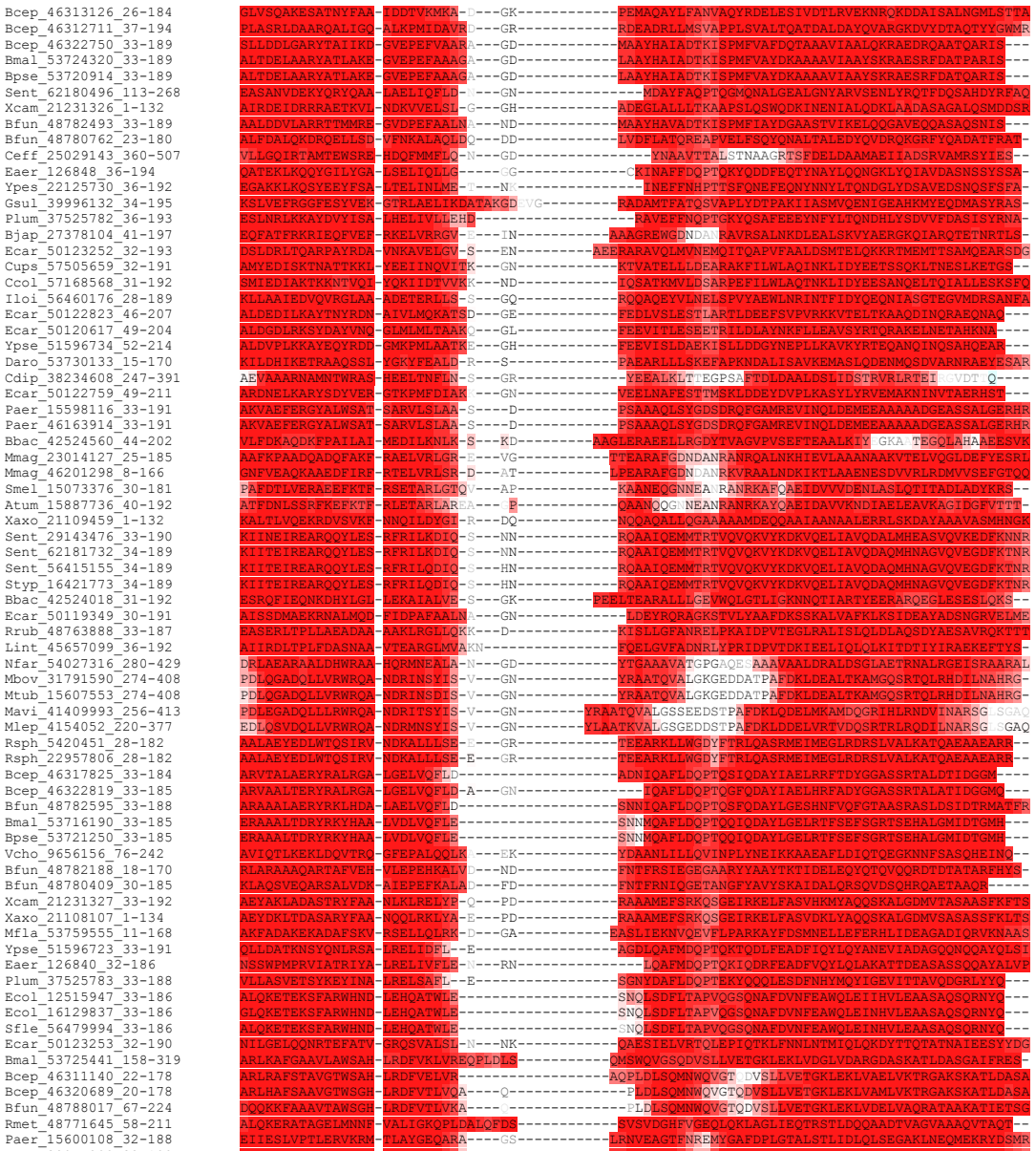


Figure B.3 continued

Figure B.3 continued

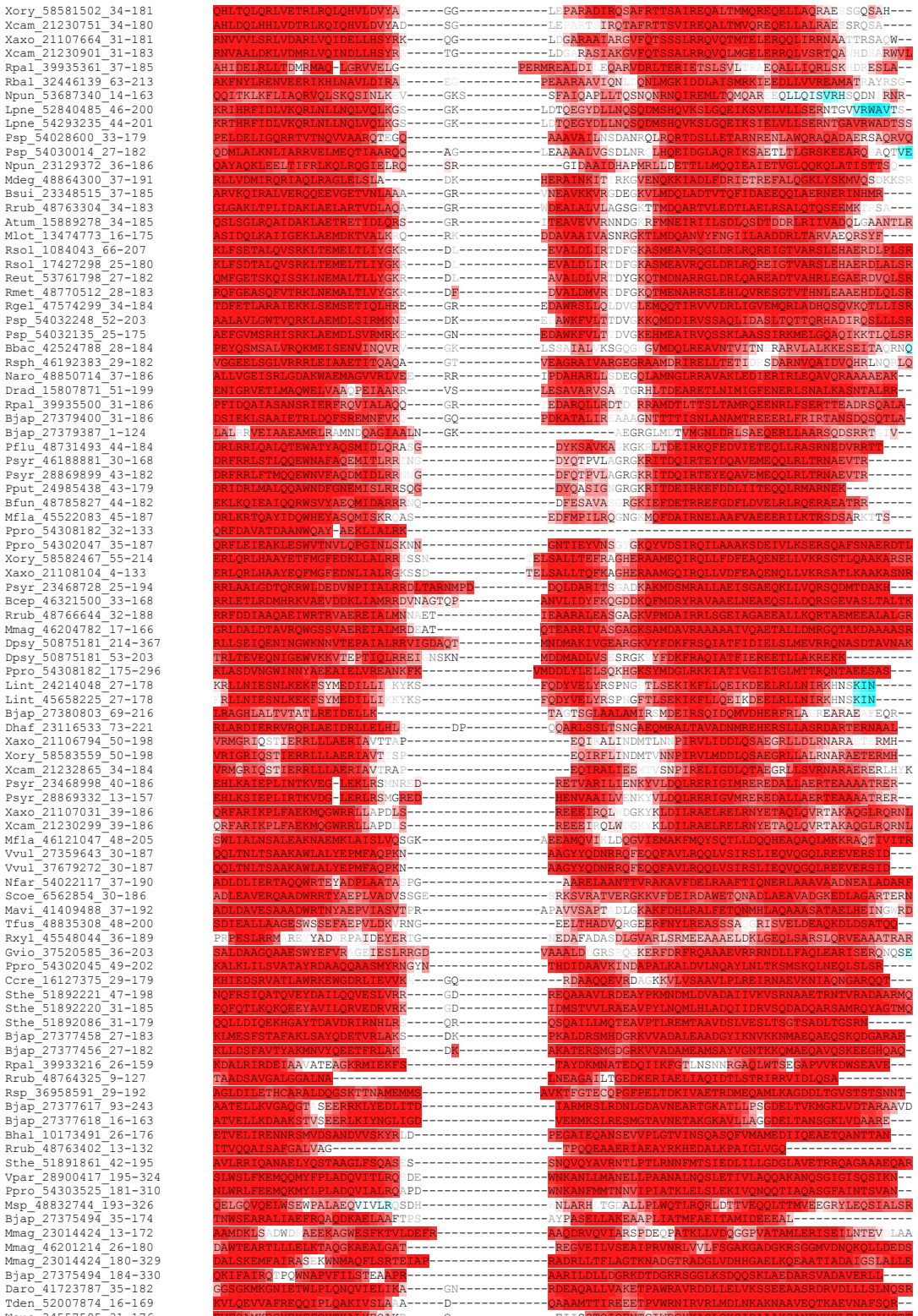
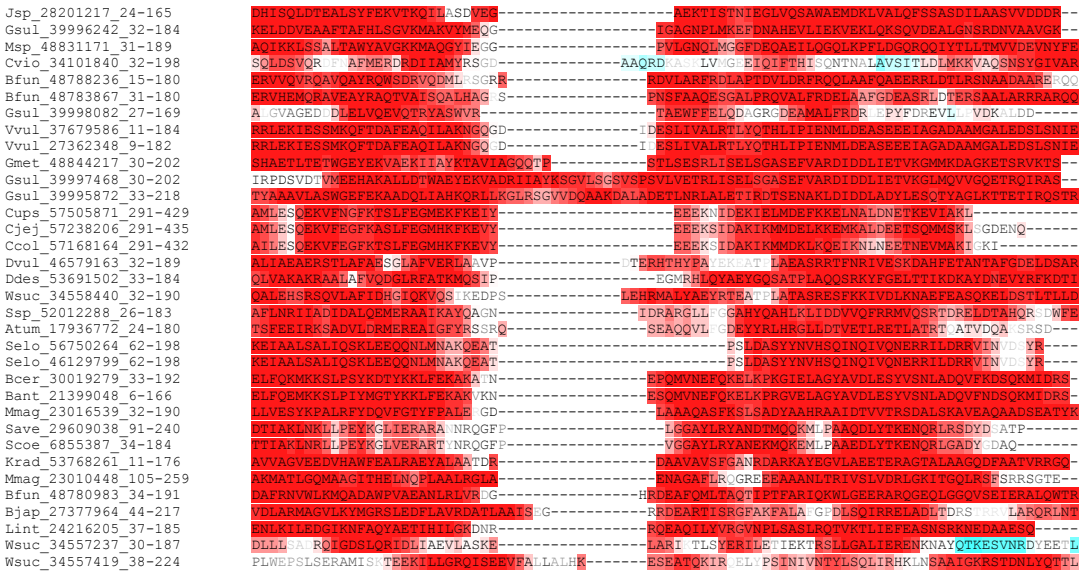


Figure B.3 continued

Figure B.3 continued



Agam, *Anopheles gambiae*; Aper, *Aeropyrum pernix*; Asp, *Azoarcus sp*; Atum, *Agrobacterium tumefaciens*; Avar, *Anabaena variabilis*; Avin, *Azotobacter vinelandii*; Bagr, *Brevibacillus agri*; Bant, *Bacillus anthracis*; Bbac, *Bdellovibrio bacteriovorus*; Bbro, *Bordetella bronchiseptica*; Bcep, *Burkholderia cepacia*; Bcer, *Bacillus cereus*; Bfun, *Burkholderia fungorum*; Bhal, *Bacillus halodurans*; Bjap, *Bradyrhizobium japonicum*; Blic, *Bacillus licheniformis*; Bmal, *Burkholderia mallei*; Bpar, *Bordetella parapertussis*; Bper, *Bordetella pertussis*; Bpse, *Burkholderia pseudomallei*; Bsub, *Bacillus subtilis*; Bsui, *Brucella suis*; Bthu, *Bacillus thuringiensis*; Cace, *Clostridium acetobutylicum*; Caur, *Chloroflexus aurantiacus*; Ccol, *Campylobacter coli*; Ccre, *Caulobacter crescentus*; Cdip, *Corynebacterium diphtheriae*; Ceff, *Corynebacterium efficiens*; Cglu, *Corynebacterium glutamicum*; Chut, *Cytophaga hutchinsonii*; Cjej, *Campylobacter jejuni*; Ctep, *Chlorobium tepidum*; Ctet, *Clostridium tetani*; Cthe, *Clostridium thermocellum*; Cups, *Campylobacter upsaliensis*; Cvio, *Chromobacterium violaceum*; Daro, *Dechloromonas aromatica*; Ddes, *Desulfovibrio desulfuricans*; Dgig, *Desulfovibrio gigas*; Dhaf, *Desulfitobacterium hafniense*; Dpsy, *Desulfotalea psychrophila*; Drad, *Deinococcus radiodurans*; Dvul, *Desulfovibrio vulgaris*; Eaer, *Enterobacter aerogenes*; Eamy, *Erwinia amylovora*; Ecar, *Erwinia carotovora*; Echr, *Erwinia chrysanthemi*; Ecol, *Escherichia coli*; Epyr, *Erwinia pyrifoliae*; Esp, *Erwinia sp*; Gkau, *Geobacillus kaustophilus*; Gmet, *Geobacter metallireducens*; Gsul, *Geobacter sulfurreducens*; Gvio, *Gloeobacter violaceus*; Hhal, *Halomonas halodenitrificans*; Hhep, *Helicobacter hepaticus*; Iloi, *Idiomarina loihiensis*; Jsp, *Janthinobacterium sp*; Krad, *Kineococcus radiotolerans*; Linn, *Listeria innocua*; Lint, *Leptospira interrogans*; Lmon, *Listeria monocytogenes*; Lpne, *Legionella pneumophila*; Mace, *Methanosarcina acetivorans*; Mavi, *Mycobacterium avium*; Mbar, *Methanosarcina barkeri*; Mbov, *Mycobacterium bovis*; Mdeg, *Microbulbifer degradans*; Mfla, *Methylobacillus flagellatus*; Mgal, *Mycoplasma gallisepticum*; Mlep, *Mycobacterium leprae*; Mlot, *Mesorhizobium loti*; Mmag, *Magnetospirillum magnetotacticum*; Mmaz, *Methanosarcina mazei*; Msp, *Magnetococcus sp*; Mtub, *Mycobacterium tuberculosis*; Naro, *Novosphingobium aromaticivorans*; Neur, *Nitrosomonas europaea*; Nfar, *Nocardia farcinica*; Npun, *Nostoc punctiforme*; Nsp, *Nostoc sp*; Paer, *Pseudomonas aeruginosa*; Pflu, *Pseudomonas fluorescens*; Plum, *Photorhabdus luminescens*; Pmul, *Pasteurella multocida*; Ppro, *Photobacterium profundum*; Pput, *Pseudomonas putida*; Pres, *Pseudomonas resinovorans*; Psp, *Polaromonas sp*; Psyr, *Pseudomonas syringae*; Raqu, *Rahnella aquatilis*; Rbal, *Rhodopirellula baltica*; Rcen, *Rhodospirillum centenum*; Retl, *Rhizobium etli*; Reut, *Ralstonia eutropha*; Rgel, *Rubrivivax gelatinosus*; Rleg, *Rhizobium leguminosarum*; Rmet, *Ralstonia metallidurans*; Rpal, *Rhodopseudomonas palustris*; Rrub, *Rhodospirillum rubrum*; Rsol, *Ralstonia solanacearum*; Rsp, *Rhizobium sp*; Rsph, *Rhodobacter sphaeroides*; Rxyl, *Rubrobacter xylanophilus*; Save, *Streptomyces avermitilis*; Scoe, *Streptomyces coelicolor*; Selo, *Synechococcus elongatus*; Sent, *Salmonella enterica*; Sfle, *Shigella flexneri*; Smel, *Sinorhizobium meliloti*; Sone, *Shewanella oneidensis*; Spla, *Spirulina platensis*; Ssp, *Silicibacter sp*; Sthe, *Symbiobacterium thermophilum*; Styp, *Salmonella typhimurium*; Tden, *Thiobacillus denitrificans*; Tfus, *Thermobifida fusca*; Vcho, *Vibrio cholerae*; Vfis, *Vibrio fischeri*; Vpar, *Vibrio parahaemolyticus*; Vvul, *Vibrio vulnificus*; Wsuc, *Wolinella succinogenes*; Xaxo, *Xanthomonas axonopodis*; Xcam, *Xanthomonas campestris*; Xory, *Xanthomonas oryzae*; Ypes, *Yersinia pestis*; Ypse, *Yersinia pseudotuberculosis*; Zmob, *Zymomonas mobilis*

Figure B.3 continued

APPENDIX C

SUPPLEMENTARY INFORMATION FOR CHAPTER 6

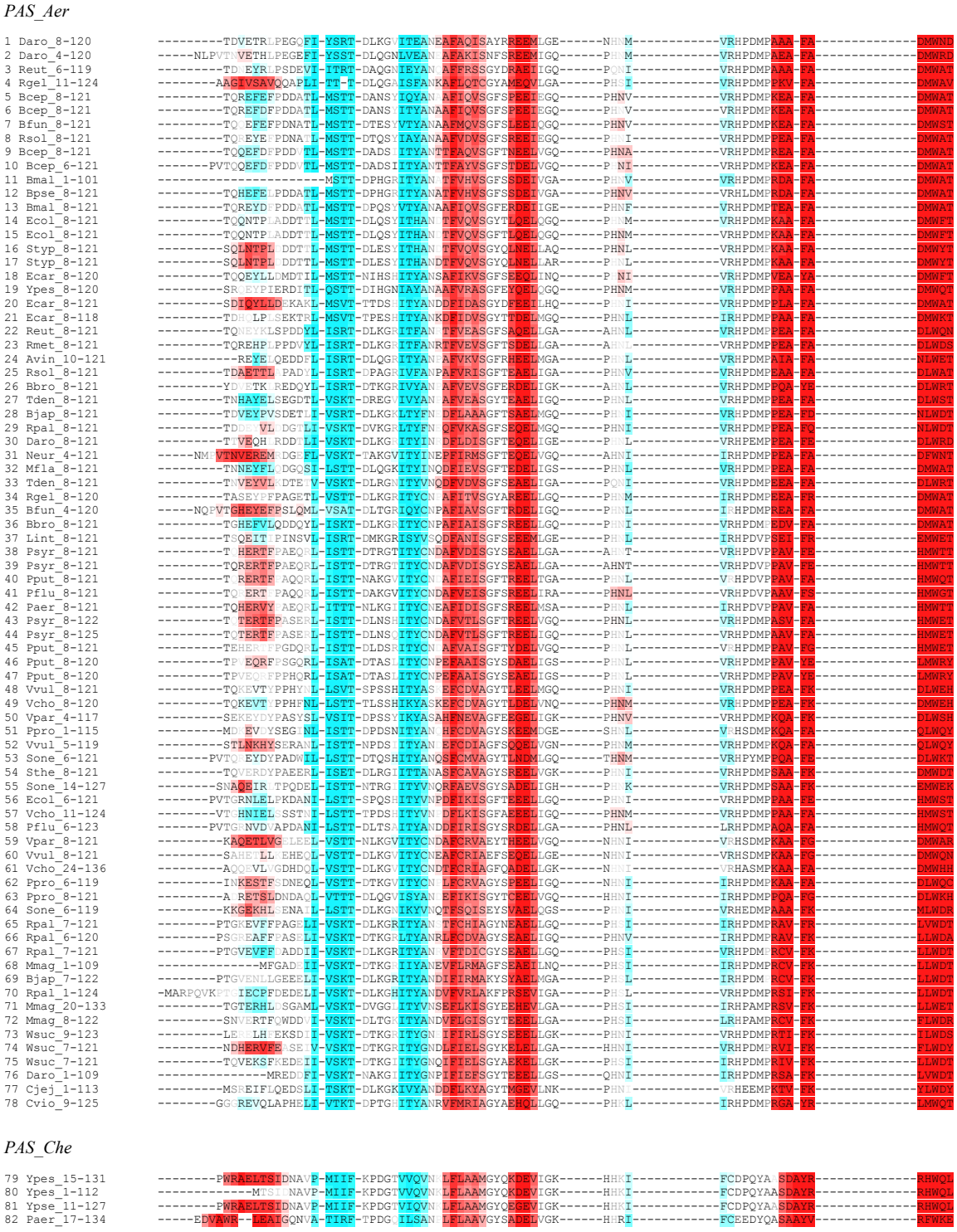


Figure C.1 Unedited alignment of chemotaxis PAS domains constructed using PCMA and ClustalW, and visualized with VISSA.

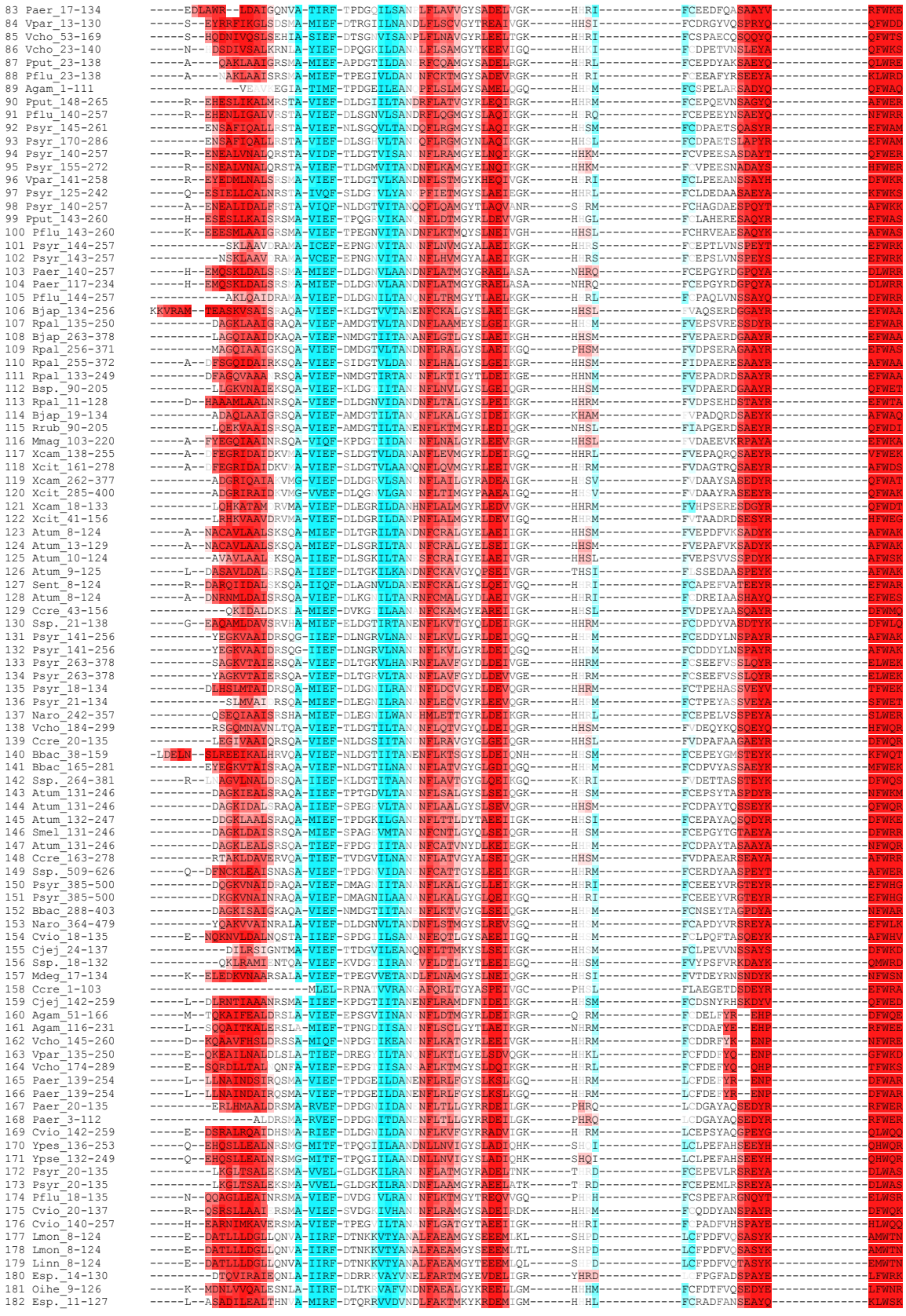
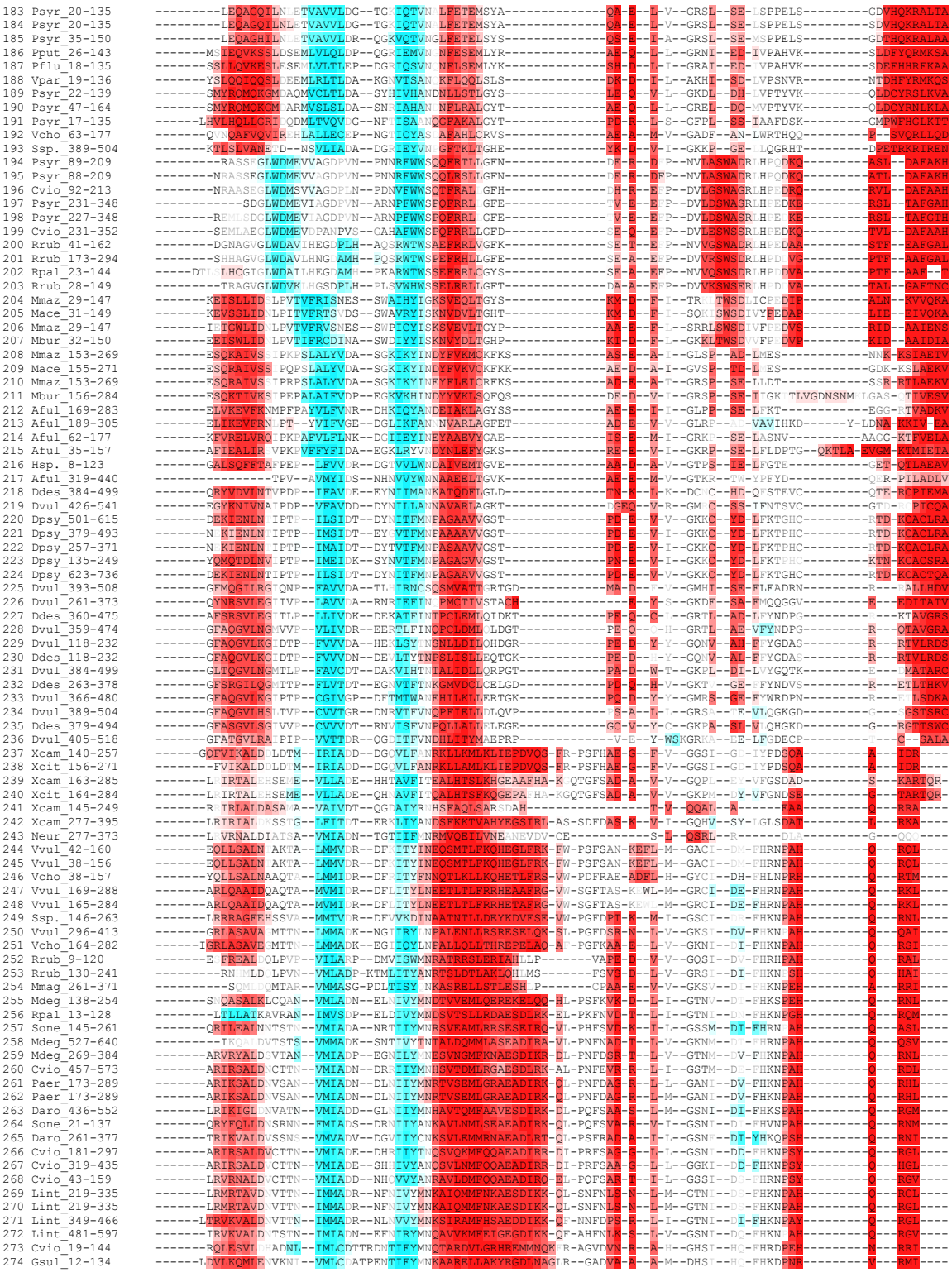


Figure C.1 continued

Generic PAS



PAS_Aer

Figure C.1 continued

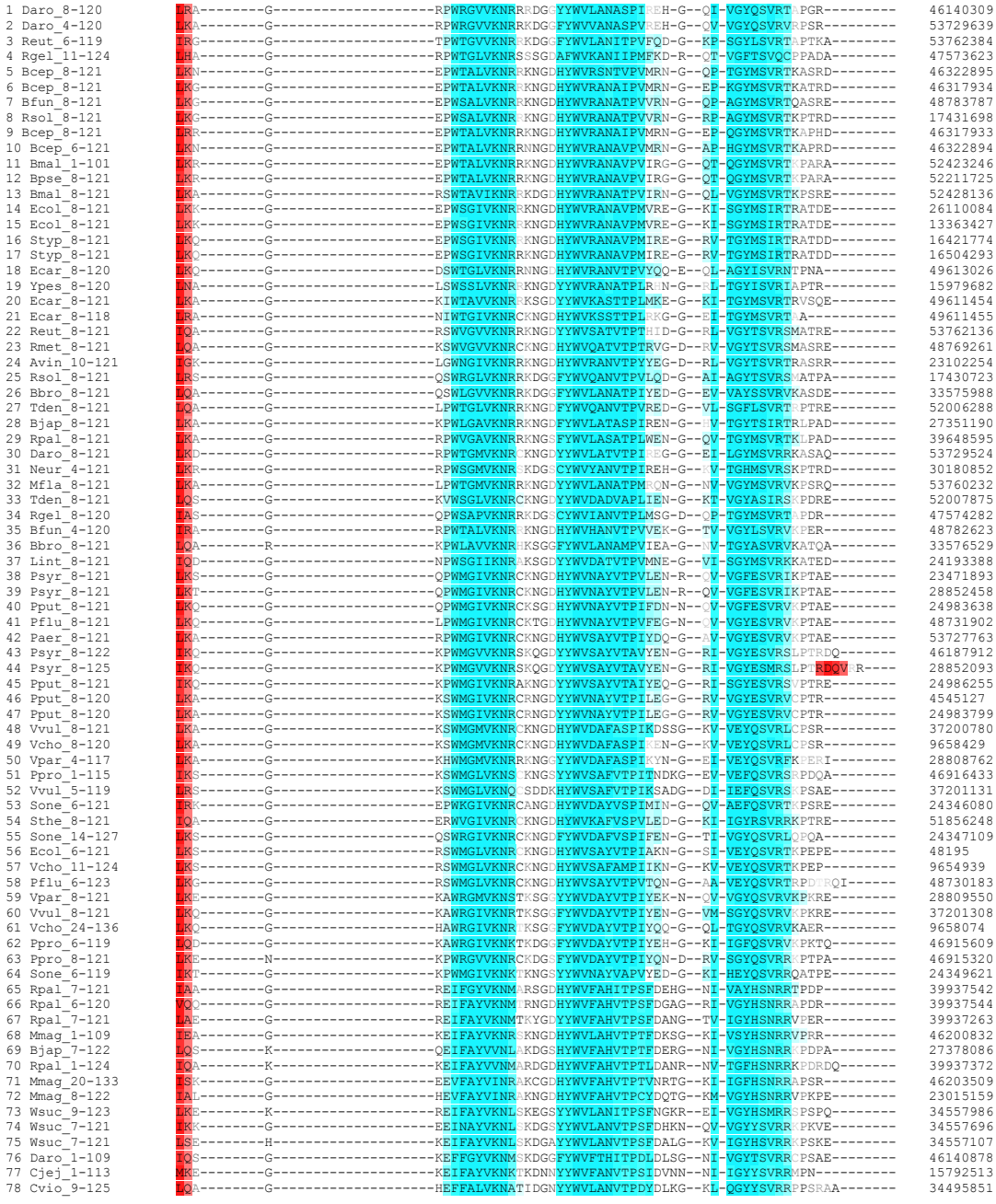
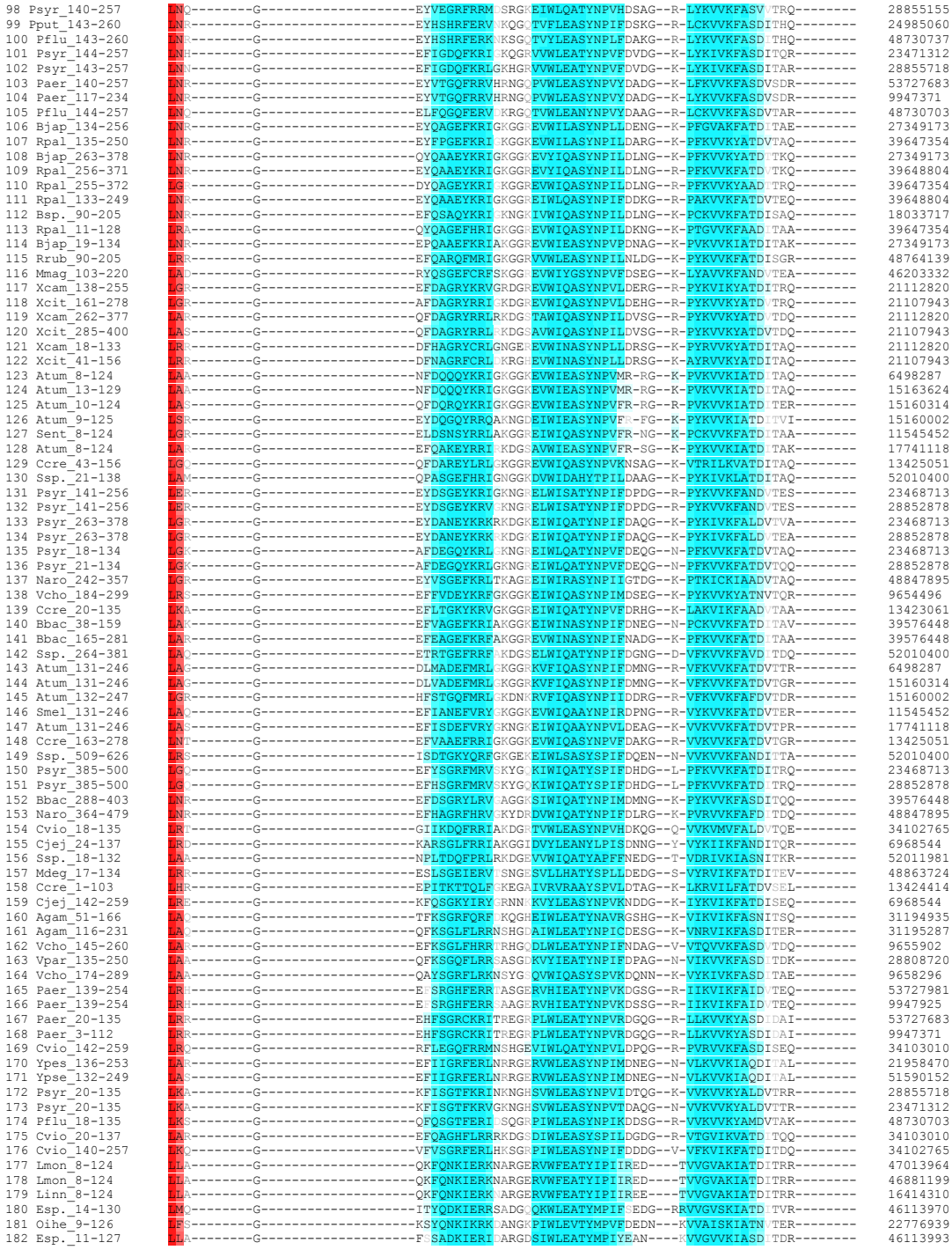


Figure C.1 continued



Generic PAS

Figure C.1 continued

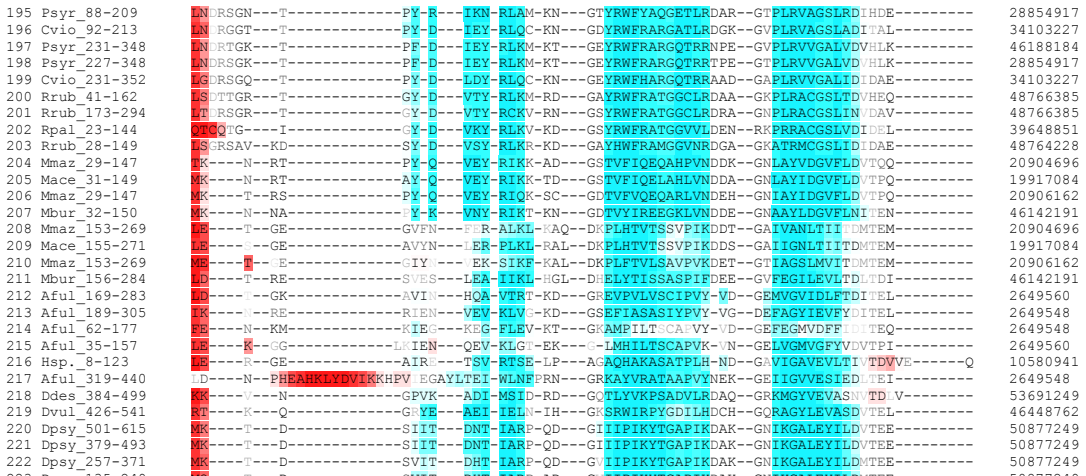
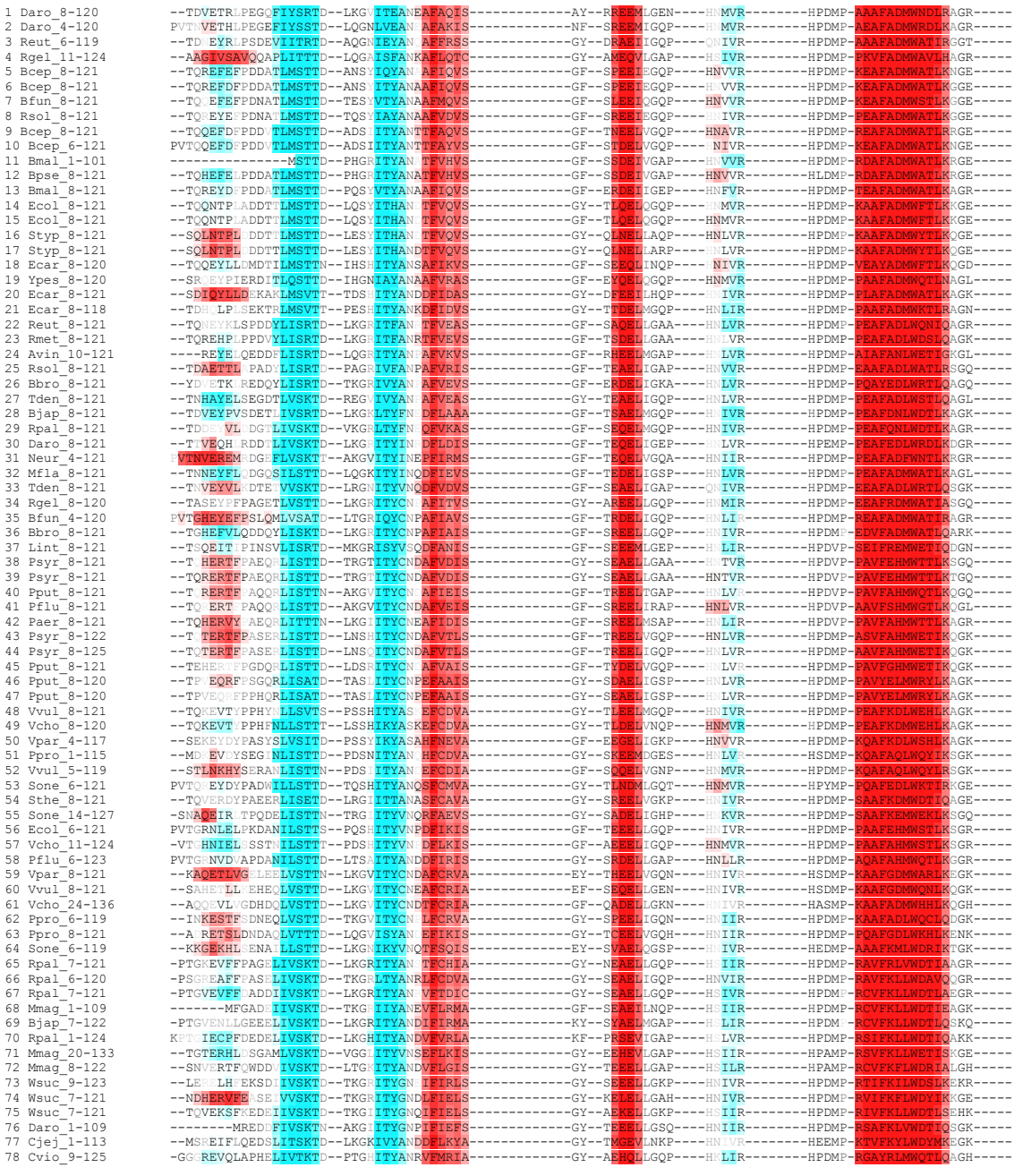


Figure C.1 continued

PAS_Aer



PAS_Che

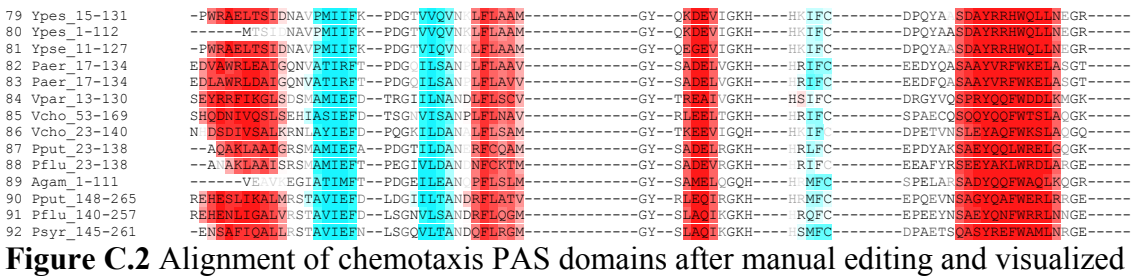
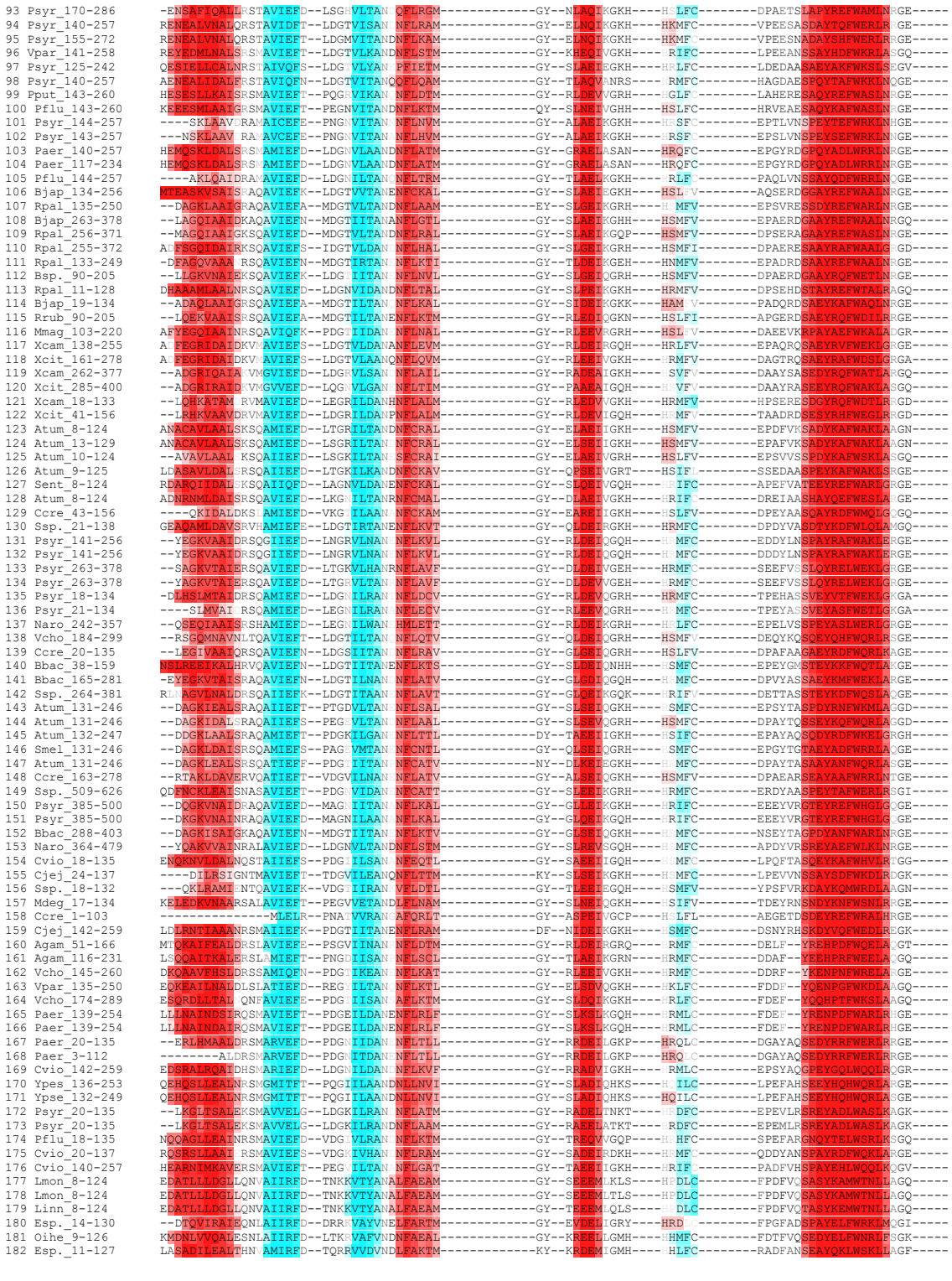


Figure C.2 Alignment of chemotaxis PAS domains after manual editing and visualized with VISSA.



Generic PAS

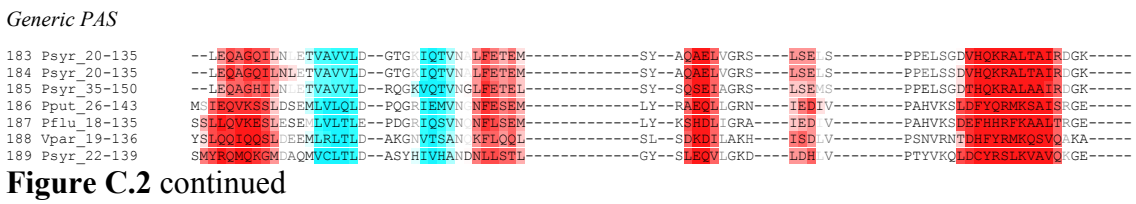
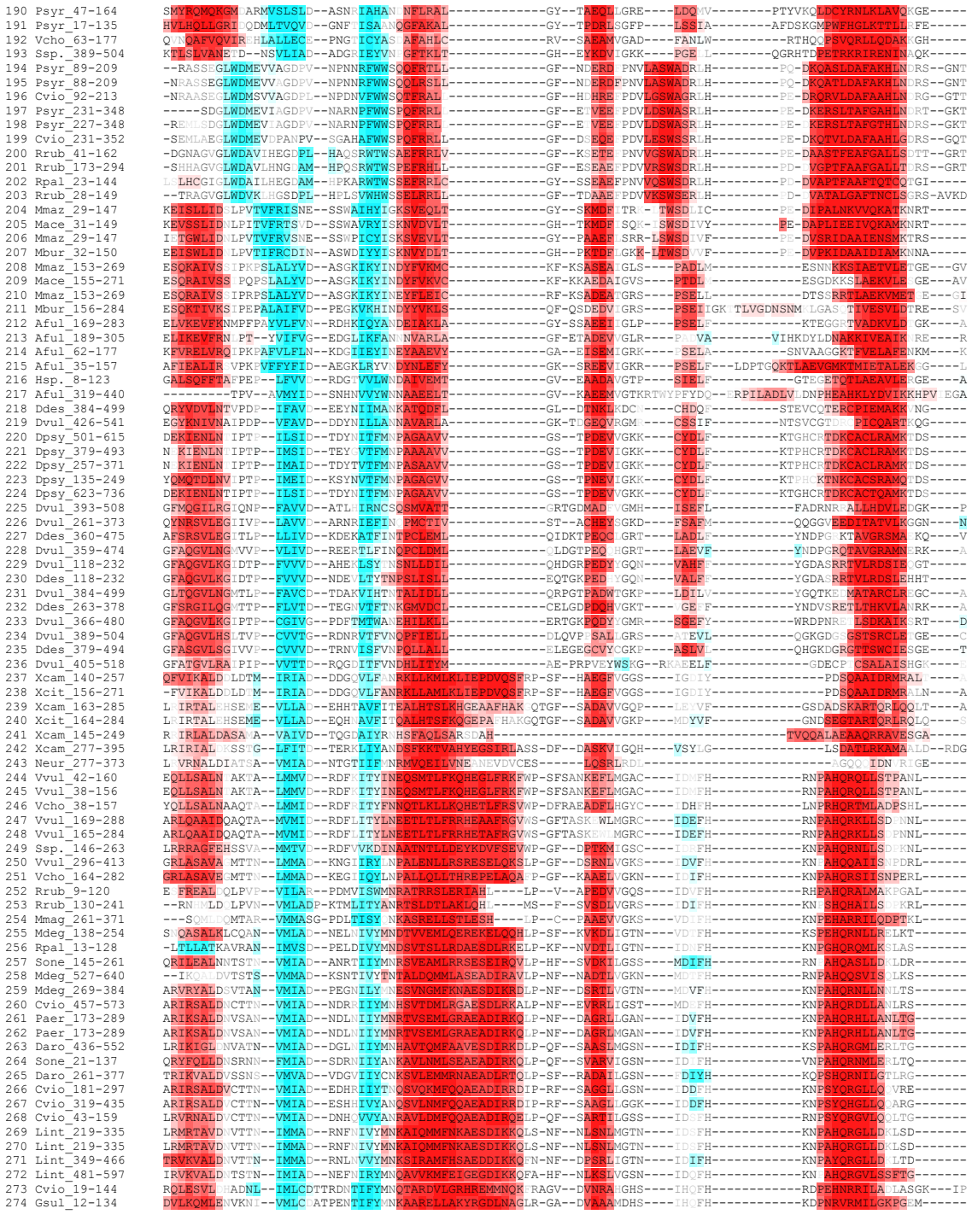


Figure C.2 continued



PAS_Aer

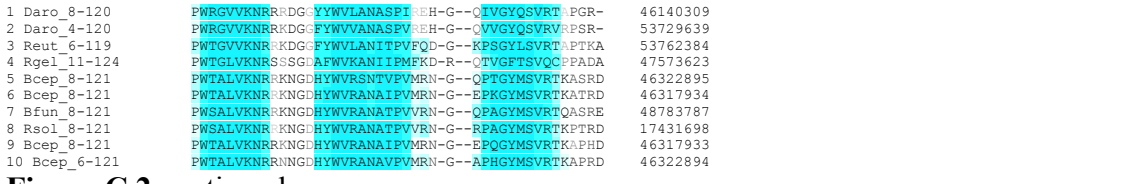
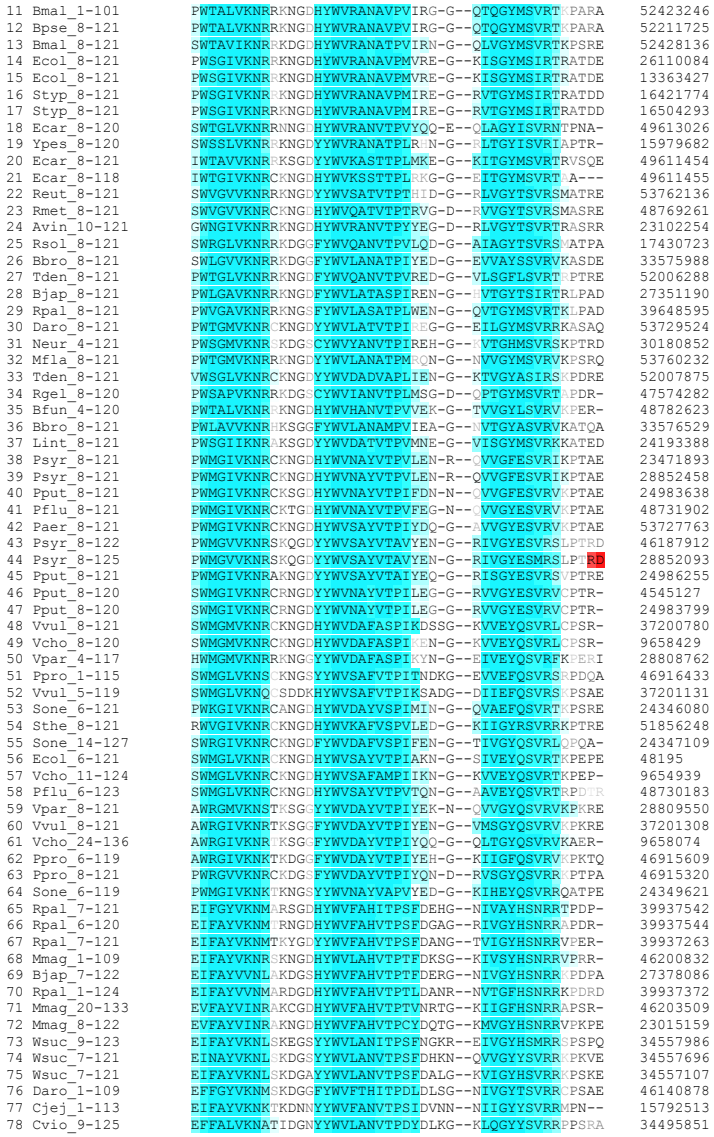


Figure C.2 continued



PAS_Che

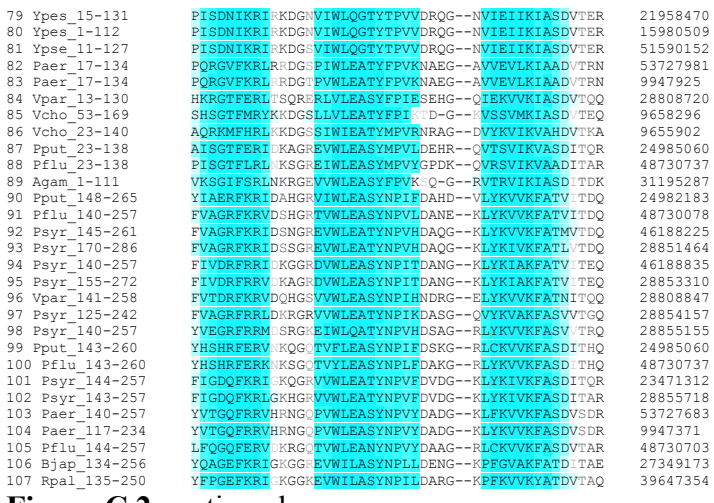
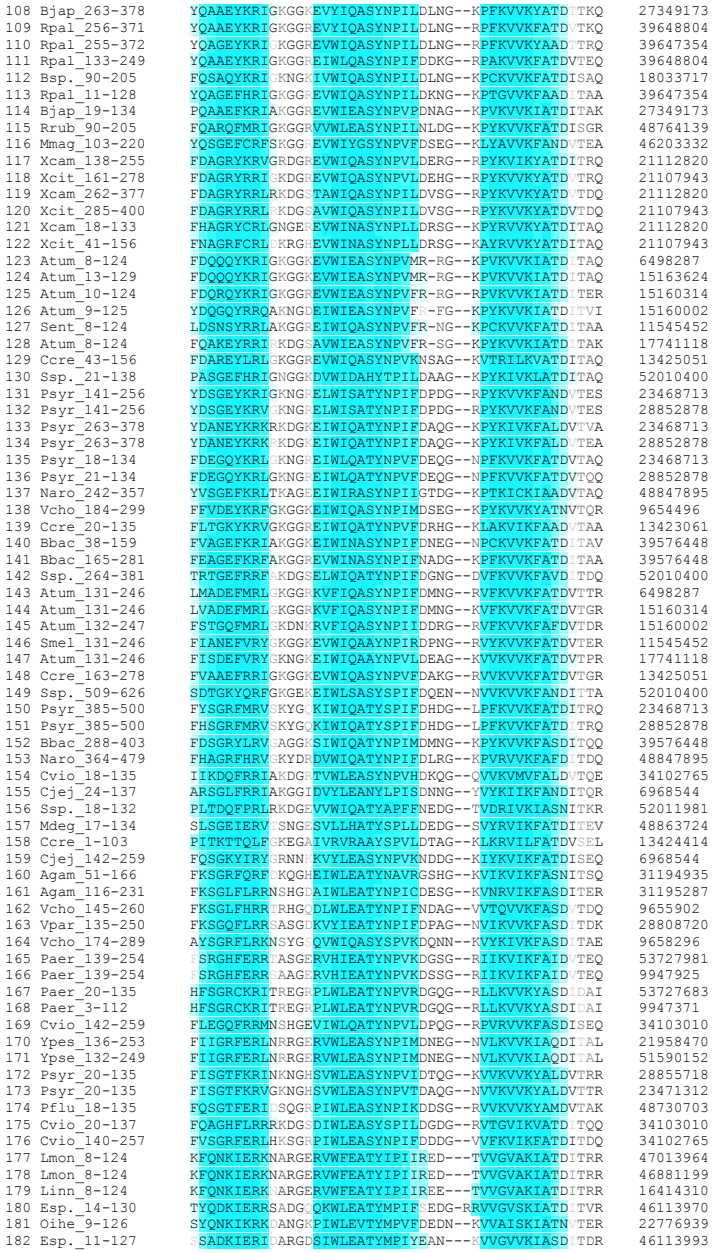


Figure C.2 continued



Generic PAS

Figure C.2 continued

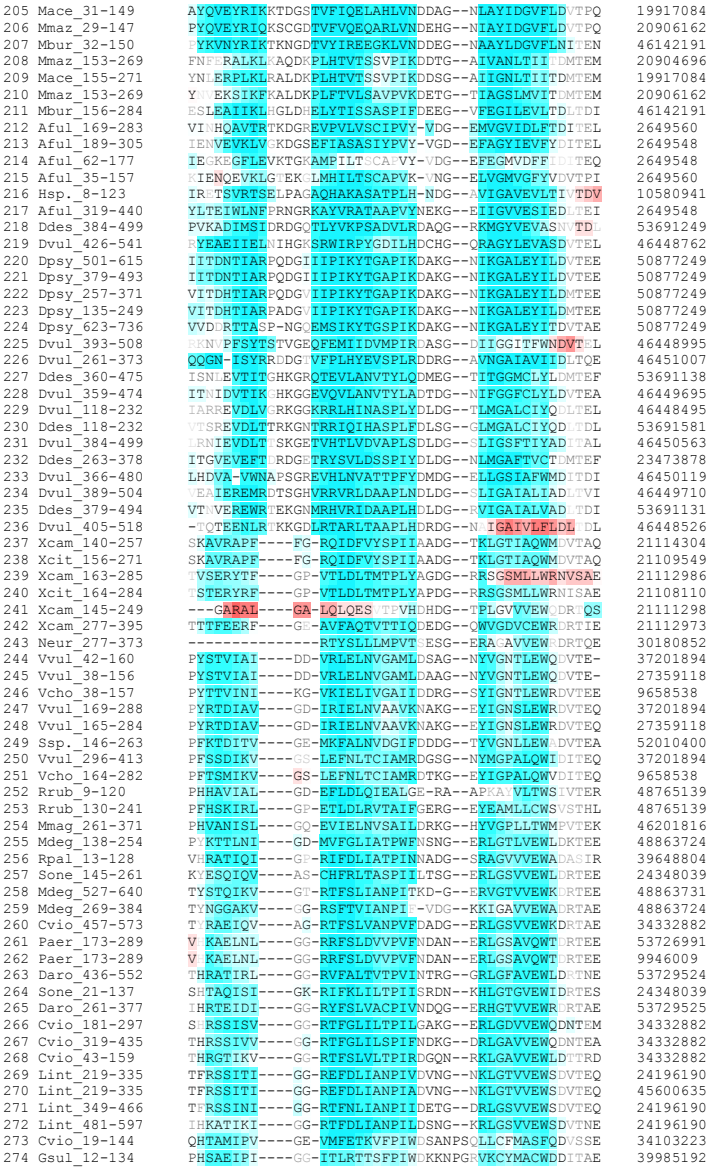


Figure C.2 continued

APPENDIX D

PUBLICATIONS

1. Shu, C.J., Ulrich, L.E., and Zhulin, I.B. (2003) The NIT domain: a predicted nitrate responsive module in bacterial sensory receptors.

2. Ulrich, L.E., Koonin, E.V., and Zhulin, I.B. (2005) One-component systems dominate signal transduction in prokaryotes, *Trends Microbiol*, **13**, 52-56.

3. Ulrich, L.E., and Zhulin, I.B. (2005) Four-Helix Bundle: a Ubiquitous Sensory Module in Prokaryotic Signal Transduction, *Bioinformatics*, in press.

4. Wu, M., Ren, Q., Durkin, A.S., Daugherty, S.C., Brinkac, L.M., Dodson, R.J., Madupu, R., Sullivan, S.A., Kolonay, J.F., Nelson, W., Tallon, L.J., Jones, K.M., Ulrich, L.E., Gonzalez, J.M., Zhulin, I.B., Robb, F.T. and Eisen, J.A. (2005) Life in Hot Carbon Monoxide: the Complete Genome Sequence of Carboxydothermus hydrogenoformans Z-2901, *PLoS Genetics*, preprint.

5. Richardson, P., Chain, P., Vergez, L., Malfatti, S., Larimer, F., Hauser, L., Land, M., Lao, V., Denef, V., Konstantinidis, K., Parnell, J., Sul, W., Tsoi, T., Marx, C., Smith, D., Chai, B., Ulrich, L.E., Zhulin, I.B., Latorre Reyes, V., Agulló, L., Gómez, L., Córdova, M., González, M., Seeger, M., Cook, A.M., LiPuma, J. (2005) The complete genome sequence of *Burkholderia xenovorans* LB400, a potent degrader of polychlorinated biphenyls, *PNAS*, submitted.

6. Ulrich, L.E., Black, W., Ma, Q., Taylor, B.L., and Zhulin, I.B. Resolving the Function of Chemotaxis PAS Domains through Protein Sequence Analysis, manuscript in preparation.

REFERENCES

Altschul, S.F., Boguski, M.S., Gish, W. and Wootton, J.C. (1994) Issues in Searching Molecular Sequence Databases, *Nat Genet*, **6**, 119-129.

Altschul, S.F. and Gish, W. (1996) Local alignment statistics, *Method Enzymol*, **266**, 460-480.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic Local Alignment Search Tool, *J Mol Biol*, **215**, 403-410.

Altschul, S.F. and Koonin, E.V. (1998) Iterated profile searches with PSI-BLAST - a tool for discovery in protein databases, *Trends Biochem Sci*, **23**, 444-447.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J.H., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res*, **25**, 3389-3402.

Anantharaman, V. and Aravind, L. (2000) Cache - a signaling domain common to animal Ca2+ channel subunits and a class of prokaryotic chemotaxis receptors, *Trends Biochem Sci*, **25**, 535-537.

Anantharaman, V. and Aravind, L. (2001) The CHASE domain: a predicted ligand-binding module in plant cytokinin receptors and other eukaryotic and bacterial receptors, *Trends Biochem Sci*, **26**, 579-582.

Anantharaman, V., Koonin, E.V. and Aravind, L. (2001) Regulatory potential, phyletic distribution and evolution of ancient, intracellular small-molecule-binding domains, *J Mol Biol*, **307**, 1271-1292.

Antoine, R., Huvent, I., Chemlal, K., Deray, I., Raze, D., Locht, C. and Jacob-Dubuisson, F. (2005) The periplasmic binding protein of a tripartite tricarboxylate transporter is involved in signal transduction, *J Mol Biol*, **351**, 799-809.

Appleby, J.L., Parkinson, J.S. and Bourret, R.B. (1996) Signal transduction via the multi-step phosphorelay: Not necessarily a road less traveled, *Cell*, **86**, 845-848.

Aravind, L. and Koonin, E.V. (1998) The HD domain defines a new superfamily of metal-dependent phosphohydrolases, *Trends Biochem Sci*, **23**, 469-472.

Aravind, L. and Ponting, C.P. (1997) The GAF domain: an evolutionary link between diverse phototransducing proteins, *Trends Biochem Sci*, **22**, 458-459.

Aravind, L. and Ponting, C.P. (1999) The cytoplasmic helical linker domain of receptor histidine kinase and methyl-accepting proteins is common to many prokaryotic signalling proteins, *Fems Microbiol Lett*, **176**, 111-116.

Armitage, J.P. (1999) Bacterial tactic responses, *Adv Microb Physiol*, **41**, 229-289.

Balazsi, G., Barabasi, A.L. and Oltvai, Z.N. (2005) Topological units of environmental signal processing in the transcriptional regulatory network of Escherichia coli, *P Natl Acad Sci USA*, **102**, 7841-7846.

Batchelor, E., Walthers, D., Kenney, L.J. and Goulian, M. (2005) The Escherichia coli CpxA-CpxR envelope stress response system regulates expression of the Porins OmpF and OmpC, *J Bacteriol*, **187**, 5723-5731.

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L.L., Studholme, D.J., Yeats, C. and Eddy, S.R. (2004) The Pfam protein families database, *Nucleic Acids Res*, **32**, D138-D141.

Bendtsen, J.D., Nielsen, H., von Heijne, G. and Brunak, S. (2004) Improved prediction of signal peptides: SignalP 3.0, *J Mol Biol*, **340**, 783-795.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Wheeler, D.L. (2005) GenBank, *Nucleic Acids Res*, **33**, D34-D38.

Bernal, A., Ear, U. and Kyrpides, N. (2001) Genomes OnLine Database (GOLD): a monitor of genome projects world-wide, *Nucleic Acids Res*, **29**, 126-127.

Besemer, J., Lomsadze, A. and Borodovsky, M. (2001) GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions, *Nucleic Acids Res*, **29**, 2607-2618.

Biswas, I. and Scott, J.R. (2003) Identification of rocA, a positive regulator of covR expression in the group A streptococcus, *J Bacteriol*, **185**, 3081-3090.

Blue, C.E. and Mitchell, T.J. (2003) Contribution of a response regulator to the virulence of Streptococcus pneumoniae is strain dependent, *Infect Immun*, **71**, 4405-4413.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S. and Schneider, M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003, *Nucleic Acids Res*, **31**, 365-370.

Bork, P., Holm, L. and Sander, C. (1994) The Immunoglobulin Fold - Structural Classification, Sequence Patterns and Common Core, *J Mol Biol*, **242**, 309-320.

Boukhvalova, M., VanBruggen, R. and Stewart, R.C. (2002) CheA kinase and chemoreceptor interaction surfaces on CheW, *J Biol Chem*, **277**, 23596-23603.

Bowie, J.U., Pakula, A.A. and Simon, M.I. (1995) The 3-Dimensional Structure of the Aspartate Receptor from Escherichia-Coli, *Acta Crystallogr D*, **51**, 145-154.

Bru, C., Courcelle, E., Carrre, S., Beausse, Y., Dalmar, S. and Kahn, D. (2005) The ProDom database of protein domain families: more emphasis on 3D, *Nucleic Acids Res*, **33**, D212-D215.

Cai, S.J. and Inouye, M. (2002) EnvZ-OmpR interaction and osmoregulation in Escherichia coli, *J Biol Chem*, **277**, 24155-24161.

Calogero, S., Gardan, R., Glaser, P., Schweizer, J., Rapoport, G. and Debarbouille, M. (1994) Rocr, a Novel Regulatory Protein Controlling Arginine Utilization in Bacillus-Subtilis, Belongs to the Ntrc/Nifa Family of Transcriptional Activators, *J Bacteriol*, **176**, 1234-1241.

Chai, W.H. and Stewart, V. (1998) NasR, a novel RNA-binding protein, mediates nitrate-responsive transcription antitermination of the Klebsiella oxytoca M5al nasF operon leader in vitro, *J Mol Biol*, **283**, 339-351.

Chan, C., Paul, R., Samoray, D., Amiot, N.C., Giese, B., Jenal, U. and Schirmer, T. (2004) Structural basis of activity and allosteric control of diguanylate cyclase, *P Natl Acad Sci USA*, **101**, 17084-17089.

Changela, A., Chen, K., Xue, Y., Holschen, J., Outten, C.E., O'Halloran, T.V. and Mondragon, A. (2003) Molecular basis of metal-ion selectivity and zeptomolar sensitivity by CueR, *Science*, **301**, 1383-1387.

Cho, H. and Winans, S.C. (2005) VirA and VirG activate the Ti plasmid repABC operon, elevating plasmid copy number in response to wound-released chemical signals, *Proc Natl Acad Sci U S A*.

Clamp, M., Cuff, J., Searle, S.M. and Barton, G.J. (2004) The Jalview Java alignment editor, *Bioinformatics*, **20**, 426-427.

Comenge, Y., Quintiliani, R., Li, L., Dubost, L., Brouard, J.P., Hugonnet, J.E. and Arthur, M. (2003) The CroRS two-component regulatory system is required for intrinsic beta-lactam resistance in Enterococcus faecalis, *J Bacteriol*, **185**, 7184-7192.

Crooks, G.E., Hon, G., Chandonia, J.M. and Brenner, S.E. (2004) WebLogo: A sequence logo generator, *Genome Res*, **14**, 1188-1190.

Cserzo, M., Eisenhaber, F., Eisenhaber, B. and Simon, I. (2004) TM or not TM: transmembrane protein prediction with low false positive rate using DAS-TMfilter, *Bioinformatics*, **20**, 136-137.

Dandekar, T., Snel, B., Huynen, M. and Bork, P. (1998) Conservation of gene order: a fingerprint of proteins that physically interact, *Trends Biochem Sci*, **23**, 324-328.

Danhorn, T., Hentzer, M., Givskov, M., Parsek, M.R. and Fuqua, C. (2004) Phosphorus limitation enhances biofilm formation of the plant pathogen Agrobacterium tumefaciens through the PhoR-PhoB regulatory system, *J Bacteriol*, **186**, 4492-4501.

Dayhoff, M.O., Schwartz, R.M. and Orcutt, B.C. (1978) A model of evolutionary change in proteins, *Atlas of Protein Sequence and Structure*, **5**, 345-352.

Delcher, A.L., Harmon, D., Kasif, S., White, O. and Salzberg, S.L. (1999) Improved microbial gene identification with GLIMMER, *Nucleic Acids Res*, **27**, 4636-4641.

Deng, W., Burland, V., Plunkett, G., Boutin, A., Mayhew, G.F., Liss, P., Perna, N.T., Rose, D.J., Mau, B., Zhou, S.G., Schwartz, D.C., Fetherston, J.D., Lindler, L.E.,

Brubaker, R.R., Plano, G.V., Straley, S.C., McDonough, K.A., Nilles, M.L., Matson, J.S., Blattner, F.R. and Perry, R.D. (2002) Genome sequence of Yersinia pestis KIM, *J Bacteriol*, **184**, 4601-4611.

Deshpande, N., Addess, K.J., Bluhm, W.F., Merino-Ott, J.C., Townsend-Merino, W., Zhang, Q., Knezevich, C., Xie, L., Chen, L., Feng, Z.K., Green, R.K., Flippen-Anderson, J.L., Westbrook, J., Berman, H.M. and Bourne, P.E. (2005) The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema, *Nucleic Acids Res*, **33**, D233-D237.

Djordjevic, S. and Stock, A.M. (1998) Structural analysis of bacterial chemotaxis proteins: Components of a dynamic signaling system, *J Struct Biol*, **124**, 189-200.

Do, C.B., Mahabhashyam, M.S.P., Brudno, M. and Batzoglou, S. (2005) ProbCons: Probabilistic consistency-based multiple sequence alignment, *Genome Res*, **15**, 330-340.

Eddy, S.R. (1998) Profile hidden Markov models, *Bioinformatics*, **14**, 755-763.

Edgar, R.C. (2004) Local homology recognition and distance measures in linear time using compressed amino acid alphabets, *Nucleic Acids Res*, **32**, 380-385.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput, *Nucleic Acids Res*, **32**, 1792-1797.

Eguchi, Y. and Utsumi, R. (2005) A novel mechanism for connecting bacterial two-component signal-transduction systems, *Trends Biochem Sci*, **30**, 70-72.

Elsen, S., Swem, L.R., Swem, D.L. and Bauer, C.E. (2004) RegB/RegA, a highly conserved redox-responding global two-component regulatory system, *Microbiol Mol Biol R*, **68**, 263-+.

Falke, J.J. and Hazelbauer, G.L. (2001) Transmembrane signaling in bacterial chemoreceptors, *Trends Biochem Sci*, **26**, 257-265.

Feng, D.F., Johnson, M.S. and Doolittle, R.F. (1985) Aligning Amino-Acid Sequences - Comparison of Commonly Used Methods, *J Mol Evol*, **21**, 112-125.

Fraser, C.M., Eisen, J.A. and Salzberg, S.L. (2000) Microbial genome sequencing, *Nature*, **406**, 799-803.

Galibert, F., Finan, T.M., Long, S.R., Puhler, A., Abola, P., Ampe, F., Barloy-Hubler, F., Barnett, M.J., Becker, A., Boistard, P., Bothe, G., Boutry, M., Bowser, L., Buhrmester, J., Cadieu, E., Capela, D., Chain, P., Cowie, A., Davis, R.W., Dreano, S., Federspiel, N.A., Fisher, R.F., Gloux, S., Godrie, T., Goffeau, A., Golding, B., Gouzy, J., Gurjal, M., Hernandez-Lucas, I., Hong, A., Huizar, L., Hyman, R.W., Jones, T., Kahn, D., Kahn, M.L., Kalman, S., Keating, D.H., Kiss, E., Komp, C., Lalaure, V., Masuy, D., Palm, C., Peck, M.C., Pohl, T.M., Portetelle, D., Purnelle, B., Ramsperger, U., Surzycki, R., Thebault, P., Vandenbol, M., Vorholter, F.J., Weidner, S., Wells, D.H., Wong, K., Yeh, K.C. and Batut, J. (2001) The composite genome of the legume symbiont Sinorhizobium meliloti, *Science*, **293**, 668-672.

Galperin, M.Y. (2005) The Molecular Biology Database Collection: 2005 update, *Nucleic Acids Res*, **33**, D5-D24.

Galperin, M.Y., Nikolskaya, A.N. and Koonin, E.V. (2001) Novel domains of the prokaryotic two-component signal transduction systems, *Fems Microbiol Lett*, **203**, 11-21.

Genick, U.K., Soltis, S.M., Kuhn, P., Canestrelli, I.L. and Getzoff, E.D. (1998) Structure at 0.85 angstrom resolution of an early protein photocycle intermediate, *Nature*, **392**, 206-209.

Georgellis, D., Kwon, O., Lin, E.C.C., Wong, S.M. and Akerley, B.J. (2001) Redox signal transduction by the ArcB sensor kinase of Haemophilus influenzae lacking the PAS domain, *J Bacteriol*, **183**, 7206-7212.

Gong, W.M., Hao, B., Mansy, S.S., Gonzalez, G., Gilles-Gonzalez, M.A. and Chan, M.K. (1998) Structure of a biological oxygen sensor: A new mechanism for heme-driven signal transduction, *P Natl Acad Sci USA*, **95**, 15177-15182.

Grebe, T.W. and Stock, J.B. (1999) The histidine protein kinase superfamily, *Adv Microb Physiol*, **41**, 139-227.

Haldimann, A., Fisher, S.L., Daniels, L.L., Walsh, C.T. and Wanner, B.L. (1997) Transcriptional regulation of the Enterococcus faecium BM4147 vancomycin resistance gene cluster by the VanS-VanR two-component regulatory system in Escherichia coli K-12, *J Bacteriol*, **179**, 5903-5913.

Harper, S.M., Neil, L.C. and Gardner, K.H. (2003) Structural basis of a phototropin light switch, *Science*, **301**, 1541-1544.

Hefti, M.H., Francoijs, K.J., de Vries, S.C., Dixon, R. and Vervoort, J. (2004) The PAS fold - A redefinition of the PAS domain based upon structural prediction, *European Journal of Biochemistry*, **271**, 1198-1208.

Hendrixson, D.R., Akerley, B.J. and DiRita, V.J. (2001) Transposon mutagenesis of Campylobacter jejuni identifies a bipartite energy taxis system required for motility, *Mol Microbiol*, **40**, 214-224.

Henikoff, S. and Henikoff, J.G. (1991) Automated Assembly of Protein Blocks for Database Searching, *Nucleic Acids Res*, **19**, 6565-6572.

Henikoff, S. and Henikoff, J.G. (1992) Amino-Acid Substitution Matrices from Protein Blocks, *P Natl Acad Sci USA*, **89**, 10915-10919.

Hoch, J.A. (2000) Two-component and phosphorelay signal transduction, *Curr Opin Microbiol*, **3**, 165-170.

Hoch, J.A. and Silhavy, T.A. (1995) *Two-component signal transduction*. ASM Press, Washington, D.C.

Inouye, M. and Dutta, R. (2003) *Histidine Kinases in Signal Transduction*. Academic Press, San Diego, CA.

Island, M.D. and Kadner, R.J. (1993) Interplay between the Membrane-Associated Uhpb and Uhpc Regulatory Proteins, *J Bacteriol*, **175**, 5028-5034.

Jenal, U. (2004) Cyclic di-guanosine-monophosphate comes of age: a novel secondary messenger involved in modulating cell surface structures in bacteria?, *Curr Opin Microbiol*, **7**, 185-191.

Jones, D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices, *J Mol Biol*, **292**, 195-202.

Kall, L., Krogh, A. and Sonnhammer, E.L.L. (2004) A combined transmembrane topology and signal peptide prediction method, *J Mol Biol*, **338**, 1027-1036.

Karlin, S. and Ghandour, G. (1985) Multiple-Alphabet Amino-Acid-Sequence Comparisons of the Immunoglobulin K-Chain Constant Domain, *P Natl Acad Sci USA*, **82**, 8597-8601.

Karp, P.D. (2001) Pathway databases: A case study in computational symbolic theories, *Science*, **293**, 2040-2044.

Katoh, K., Misawa, K., Kuma, K. and Miyata, T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform, *Nucleic Acids Res*, **30**, 3059-3066.

Keseler, I.M., Collado-Vides, J., Gama-Castro, S., Ingraham, J., Paley, S., Paulsen, I.T., Peralta-Gill, M. and Karp, P.D. (2005) EcoCyc: a comprehensive database resource for Escherichia coli, *Nucleic Acids Res*, **33**, D334-D337.

Kimura, M. (1985) *The Neutral Theory of Molecular Evolution*. Cambridge University Press.

Koh, I.Y.Y., Eyrich, V.A., Marti-Renom, M.A., Przybylski, D., Madhusudhan, M.S., Eswar, N., Grana, O., Pazos, F., Valencia, A., Sali, A. and Rost, B. (2003) EVA: evaluation of protein structure prediction servers, *Nucleic Acids Res*, **31**, 3311-3315.

Kolb, A., Busby, S., Buc, H., Garges, S. and Adhya, S. (1993) Transcriptional Regulation by Camp and Its Receptor Protein, *Annu Rev Biochem*, **62**, 749-795.

Konstantinidis, K.T. and Tiedje, J.M. (2004) Trends between gene content and genome size in prokaryotic species with larger genomes, *P Natl Acad Sci USA*, **101**, 3160-3165.

Koonin, E.V. and Galperin, M.Y. (2003) *Sequence - Evolution - Function : Computational Approaches in Comparative Genomics*. Kluwer Academic Publishers, Boston, MA.

Koretke, K.K., Lupas, A.N., Warren, P.V., Rosenberg, M. and Brown, J.R. (2000) Evolution of two-component signal transduction, *Mol Biol Evol*, **17**, 1956-1970.

Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E.L.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes, *J Mol Biol*, **305**, 567-580.

Krukonis, E.S. and DiRita, V.J. (2003) From motility to virulence: sensing and responding to environmental signals in Vibrio cholerae, *Curr Opin Microbiol*, **6**, 186-190.

Kumar, S., Tamura, K. and Nei, M. (2004) MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment, *Brief Bioinform*, **5**, 150-163.

Letunic, I., Copley, R.R., Schmidt, S., Ciccarelli, F.D., Doerks, T., Schultz, J., Ponting, C.P. and Bork, P. (2004) SMART 4.0: towards genomic data integration, *Nucleic Acids Res*, **32**, D142-D144.

Lewis, M., Chang, G., Horton, N.C., Kercher, M.A., Pace, H.C., Schumacher, M.A., Brennan, R.G. and Lu, P.Z. (1996) Crystal structure of the lactose operon repressor and its complexes with DNA and inducer, *Science*, **271**, 1247-1254.

Lipman, D.J., Altschul, S.F. and Kececioglu, J.D. (1989) A Tool for Multiple Sequence Alignment, *P Natl Acad Sci USA*, **86**, 4412-4415.

Lupas, A. (1997) Predicting coiled-coil regions in proteins, *Curr Opin Struc Biol*, **7**, 388-393.

Macheroux, P., Hill, S., Austin, S., Eydmann, T., Jones, T., Kim, S.O., Poole, R. and Dixon, R. (1998) Electron donation to the flavoprotein NifL, a redox-sensing transcriptional regulator, *Biochemical Journal*, **332**, 413-419.

Matsushita, M. and Janda, K.D. (2002) Histidine kinases as targets for new antimicrobial agents, *Bioorgan Med Chem*, **10**, 855-867.

McCluskey, J., Hinds, J., Husain, S., Witney, A. and Mitchell, T.J. (2004) A two-component system that controls the expression of pneumococcal surface antigen A (PsaA) and regulates virulence and resistance to oxidative stress in Streptococcus pneumoniae, *Mol Microbiol*, **51**, 1661-1675.

Milburn, M.V., Prive, G.G., Milligan, D.L., Scott, W.G., Yeh, J., Jancarik, J., Koshland, D.E. and Kim, S.H. (1991) Three-Dimensional Structures of the Ligand-Binding Domain of the Bacterial Aspartate Receptor with and without a Ligand, *Science*, **254**, 1342-1347.

Mishra, M., Parise, G., Jackson, K.D., Wozniak, D.J. and Deora, R. (2005) The BvgAS signal transduction system regulates biofilm development in Bordetella, *J Bacteriol*, **187**, 1474-1484.

Mougel, C. and Zhulin, I.B. (2001) CHASE: an extracellular sensing domain common to transmembrane receptors from prokaryotes, lower eukaryotes and plants, *Trends Biochem Sci*, **26**, 582-584.

Moult, J., Hubbard, T., Bryant, S.H., Fidelis, K. and Pedersen, J.T. (1997) Critical assessment of methods of protein structure prediction (CASP): Round II, *Proteins*, 2-6.

Needleman, S.B. and Wunsch, C.D. (1970) A General Method Applicable to Search for Similarities in Amino Acid Sequence of 2 Proteins, *J Mol Biol*, **48**, 443-&.

Nichols, N.N. and Harwood, C.S. (2000) An aerotaxis transducer gene from Pseudomonas putida, *Fems Microbiol Lett*, **182**, 177-183.

Nikolskaya, A.N. and Galperin, M.Y. (2002) A novel type of conserved DNA-binding domain in the transcriptional regulators of the AlgR/AgrA/LytR family, *Nucleic Acids Res*, **30**, 2453-2459.

Nioche, P., Berka, V., Vipond, J., Minton, N., Tsai, A.L. and Raman, C.S. (2004) Femtomolar sensitivity of a NO sensor from Clostridium botulinum, *Science*, **306**, 1550-1553.

Notredame, C., Higgins, D.G. and Heringa, J. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment, *J Mol Biol*, **302**, 205-217.

Novick, R.P. and Jiang, D.R. (2003) The staphylococcal saeRS system coordinates environmental signals with agr quorum sensing, *Microbiol-Sgm*, **149**, 2709-2717.

Ohki, R., Giyanto, Tateno, K., Masuyama, W., Moriya, S., Kobayashi, K. and Ogasawara, N. (2003) The BceRS two-component regulatory system induces

expression of the bacitracin transporter, BceAB, in Bacillus subtilis, *Mol Microbiol*, **49**, 1135-1144.

Ottemann, K.M., Xiao, W.Z., Shin, Y.K. and Koshland, D.E. (1999) A piston model for transmembrane signaling of the aspartate receptor, *Science*, **285**, 1751-1754.

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D. and Maltsev, N. (1999) The use of gene clusters to infer functional coupling, *P Natl Acad Sci USA*, **96**, 2896-2901.

Pappas, K.M., Weingart, C.L. and Winans, S.C. (2004) Chemical communication in proteobacteria: biochemical and structural studies of signal synthases and receptors required for intercellular signalling, *Mol Microbiol*, **53**, 755-769.

Parkinson, J.S. (1993) Signal-Transduction Schemes of Bacteria, *Cell*, **73**, 857-871.

Parkinson, J.S. and Kofoid, E.C. (1992) Communication Modules in Bacterial Signaling Proteins, *Annu Rev Genet*, **26**, 71-112.

Paul, R., Weiser, S., Amiot, N.C., Chan, C., Schirmer, T., Giese, B. and Jenal, U. (2004) Cell cycle-dependent dynamic localization of a bacterial response regulator with a novel di-guanylate cyclase output domain, *Gene Dev*, **18**, 715-727.

Peach, M.L., Hazelbauer, G.L. and Lybrand, T.P. (2002) Modeling the transmembrane domain of bacterial chemoreceptors, *Protein Sci*, **11**, 912-923.

Pearson, W.R. and Lipman, D.J. (1988) Improved Tools for Biological Sequence Comparison, *P Natl Acad Sci USA*, **85**, 2444-2448.

Pei, J.M. and Grishin, N.V. (2001) GGDEF domain is homologous to adenylyl cyclase, *Proteins*, **42**, 210-216.

Pei, J.M., Sadreyev, R. and Grishin, N.V. (2003) PCMA: fast and accurate multiple sequence alignment based on profile consistency, *Bioinformatics*, **19**, 427-428.

Pellequer, J.L., Brudler, R. and Getzoff, E.D. (1999) Biological sensors: More than one way to sense oxygen, *Current Biology*, **9**, R416-R418.

Pellequer, J.L., Wager-Smith, K.A., Kay, S.A. and Getzoff, E.D. (1998) Photoactive yellow protein: A structural prototype for the three-dimensional fold of the PAS domain superfamily, *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 5884-5890.

Ponting, C.P. and Aravind, L. (1997) PAS: a multifunctional domain family comes to light, *Current Biology*, **7**, R674-R677.

Pruitt, K.D., Tatusova, T. and Maglott, D.R. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins, *Nucleic Acids Res*, **33**, D501-D504.

Read, T.D., Peterson, S.N., Tourasse, N., Baillie, L.W., Paulsen, I.T., Nelson, K.E., Tettelin, H., Fouts, D.E., Eisen, J.A., Gill, S.R., Holtzapple, E.K., Okstad, O.A., Helgason, E., Rilstone, J., Wu, M., Kolonay, J.F., Beanan, M.J., Dodson, R.J., Brinkac, L.M., Gwinn, M., DeBoy, R.T., Madpu, R., Daugherty, S.C., Durkin, A.S., Haft, D.H., Nelson, W.C., Peterson, J.D., Pop, M., Khouri, H.M., Radune, D., Benton, J.L., Mahamoud, Y., Jiang, L.X., Hance, I.R., Weidman, J.F., Berry, K.J., Plaut, R.D., Wolf, A.M., Watkins, K.L., Nierman, W.C., Hazen, A., Cline, R., Redmond, C., Thwaite, J.E., White, O., Salzberg, S.L., Thomason, B., Friedlander, A.M., Koehler, T.M., Hanna, P.C., Kolsto, A.B. and Fraser, C.M. (2003) The genome sequence of Bacillus anthracis Ames and comparison to closely related bacteria, *Nature*, **423**, 81-86.

Rebbapragada, A., Johnson, M.S., Harding, G.P., Zuccarelli, A.J., Fletcher, H.M., Zhulin, I.B. and Taylor, B.L. (1997) The Aer protein and the serine chemoreceptor Tsr independently sense intracellular energy levels and transduce oxygen, redox, and energy signals for Escherichia coli behavior, *P Natl Acad Sci USA*, **94**, 10541-10546.

Reid, C.J. and Poole, P.S. (1998) Roles of DctA and DctB in signal detection by the dicarboxylic acid transport system of Rhizobium leguminosarum, *J Bacteriol*, **180**, 2660-2669.

Reinelt, S., Hofmann, E., Gerharz, T., Bott, M. and Madden, D.R. (2003) The structure of the periplasmic ligand-binding domain of the sensor kinase CitA reveals the first extracellular PAS domain, *J Biol Chem*, **278**, 39189-39196.

Repik, A., Rebbapragada, A., Johnson, M.S., Haznedar, J.O., Zhulin, I.B. and Taylor, B.L. (2000) PAS domain residues involved in signal transduction by the Aer redox sensor of Escherichia coli, *Molecular Microbiology*, **36**, 806-816.

Rogers, Y.H. and Venter, J.C. (2005) Genomics - Massively parallel sequencing, *Nature*, **437**, 326-327.

Rost, B., Sander, C. and Schneider, R. (1994) Redefining the Goals of Protein Secondary Structure Prediction, *J Mol Biol*, **235**, 13-26.

Saitou, N. and Nei, M. (1987) The Neighbor-Joining Method - a New Method for Reconstructing Phylogenetic Trees, *Mol Biol Evol*, **4**, 406-425.

Schaffer, A.A., Aravind, L., Madden, T.L., Shavirin, S., Spouge, J.L., Wolf, Y.I., Koonin, E.V. and Altschul, S.F. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements, *Nucleic Acids Res*, **29**, 2994-3005.

Schuster, M., Silversmith, R.E. and Bourret, R.B. (2001) Conformational coupling in the chemotaxis response regulator CheY, *P Natl Acad Sci USA*, **98**, 6003-6008.

Shelver, D., Kerby, R.L., He, Y.P. and Roberts, G.P. (1997) CooA, a CO-sensing transcription factor from Rhodospirillum rubrum, is a CO-binding heme protein, *P Natl Acad Sci USA*, **94**, 11216-11220.

Shu, C.J., Ulrich, L.E. and Zhulin, I.B. (2003) The NIT domain: a predicted nitrate-responsive module in bacterial sensory receptors, *Trends Biochem Sci*, **28**, 121-124.

Shu, C.Y.J. and Zhulin, I.B. (2002) ANTAR: an RNA-binding domain in transcription antitermination regulatory proteins, *Trends Biochem Sci*, **27**, 3-5.

Smith, T.F. and Waterman, M.S. (1981) Identification of Common Molecular Subsequences, *J Mol Biol*, **147**, 195-197.

Sourjik, V. (2004) Receptor clustering and signal processing in E coli chemotaxis, *Trends Microbiol*, **12**, 569-576.

Spiro, S. and Guest, J.R. (1990) Fnr and Its Role in Oxygen-Regulated Gene-Expression in Escherichia-Coli, *Fems Microbiol Rev*, **75**, 399-428.

Staal, M., Meysman, F.J.R. and Stal, L.J. (2003) Temperature excludes N2-fixing heterocystous cyanobacteria in the tropical oceans, *Nature*, **425**, 504-507.

Stephenson, K. and Hoch, J.A. (2002) Evolution of signalling in the sporulation phosphorelay, *Mol Microbiol*, **46**, 297-304.

Stephenson, K. and Hoch, J.A. (2002) Two-component and phosphorelay signal-transduction systems as therapeutic targets, *Curr Opin Pharmacol*, **2**, 507-512.

Stock, A.M., Robinson, V.L. and Goudreau, P.N. (2000) Two-component signal transduction, *Annu Rev Biochem*, **69**, 183-215.

Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., Rao, B.S., Smirnov, S., Sverdlov, A.V., Vasudevan, S., Wolf, Y.I., Yin, J.J. and Natale, D.A. (2003) The COG database: an updated version includes eukaryotes, *Bmc Bioinformatics*, **4**, -.

Tatusov, R.L., Koonin, E.V. and Lipman, D.J. (1997) A genomic perspective on protein families, *Science*, **278**, 631-637.

Taylor, B.L. and Zhulin, I.B. (1999) PAS domains: Internal sensors of oxygen, redox potential, and light, *Microbiology and Molecular Biology Reviews*, **63**, 479-506.

Taylor, B.L. and Zhulin, I.B. (1999) PAS domains: Internal sensors of oxygen, redox potential, and light, *Microbiol Mol Biol R*, **63**, 479-+.

Taylor, B.L., Zhulin, I.B. and Johnson, M.S. (1999) Aerotaxis and other energy-sensing behavior in bacteria, *Annu Rev Microbiol*, **53**, 103-128.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) Clustal-W - Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice, *Nucleic Acids Res*, **22**, 4673-4680.

Ulijasz, A.T., Andes, D.R., Glasner, J.D. and Weisblum, B. (2004) Regulation of iron transport in Streptococcus pneumoniae by RitR, an orphan response regulator, *J Bacteriol*, **186**, 8123-8136.

Ulrich, L.E., Koonin, E.V. and Zhulin, I.B. (2005) One-component systems dominate signal transduction in prokaryotes, *Trends Microbiol*, **13**, 52-56.

Ulrich, L.E. and Zhulin, I.B. (2005) Four-Helix Bundle: a Ubiquitous Sensory Module in Prokaryotic Signal Transduction, *Bioinformatics*, in press.

Van Nimwegen, E. (2003) Scaling laws in the functional content of genomes, *Trends Genet*, **19**, 479-484.

Vannini, A., Volpari, C., Gargioli, C., Muraglia, E., Cortese, R., De Francesco, R., Neddermann, P. and Di Marco, S. (2002) The crystal structure of the quorum sensing protein TraR bound to its autoinducer and target DNA, *Embo J*, **21**, 4393-4401.

von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P. and Snel, B. (2003) STRING: a database of predicted functional associations between proteins, *Nucleic Acids Res*, **31**, 258-261.

Walderhaug, M.O., Polarek, J.W., Voelkner, P., Daniel, J.M., Hesse, J.E., Altendorf, K. and Epstein, W. (1992) Kdpd and Kdpe, Proteins That Control Expression of the Kdpabc Operon, Are Members of the 2-Component Sensor-Effector Class of Regulators, *J Bacteriol*, **174**, 2152-2159.

Weiss, V., Kramer, G., Dunnebier, T. and Flotho, A. (2002) Mechanism of regulation of the bifunctional histidine kinase NtrB in Escherichia coli, *J Mol Microb Biotech*, **4**, 229-233.

Wootton, J.C. and Federhen, S. (1993) Statistics of Local Complexity in Amino-Acid-Sequences and Sequence Databases, *Comput Chem*, **17**, 149-163.

Zahrt, T.C. and Deretic, V. (2001) Mycobacterium tuberculosis signal transduction system required for persistent infections, *P Natl Acad Sci USA*, **98**, 12706-12711.

Zhang, H.Z., Hackbarth, C.J., Chansky, K.M. and Chambers, H.F. (2001) A proteolytic transmembrane signaling pathway and resistance to beta-lactams in staphylococci, *Science*, **291**, 1962-1965.

Zhao, R., Collins, E.J., Bourret, R.B. and Silversmith, R.E. (2002) Structure and catalytic mechanism of the E-coli chernotaxis phosphatase CheZ, *Nat Struct Biol*, **9**, 570-575.

Zhou, H. and Zhou, Y. (2005) SPEM: improving multiple sequence alignment with sequence profiles and predicted secondary structures, *Bioinformatics*, **21**, 3615-3621.

Zhulin, I.B. (2001) The superfamily of chemotaxis transducers: From physiology to genomics and back, *Advances in Microbial Physiology, Vol 45*, **45**, 157-198.

Zhulin, I.B., Nikolskaya, A.N. and Galperin, M.Y. (2003) Common extracellular sensory domains in transmembrane receptors for diverse signal transduction pathways in Bacteria and Archaea, *J Bacteriol*, **185**, 285-294.

Zhulin, I.B. and Taylor, B.L. (1997) PAS domain S-boxes in Archaea, bacteria and sensors for oxygen and redox, *Trends in Biochemical Sciences*, **22**, 331-333.

Zimmer, D.P., Soupene, E., Lee, H.L., Wendisch, V.F., Khodursky, A.B., Peter, B.J., Bender, R.A. and Kustu, S. (2000) Nitrogen regulatory protein C-controlled genes of Escherichia coli: Scavenging as a defense against nitrogen limitation, *P Natl Acad Sci USA*, **97**, 14674-14679.